

AD-A174 759

WORK PERFORMANCE RATINGS: A META-ANALYSIS OF

1/2

MULTITRAIT-MULTIMETHOD STUDIES(U) TEXAS MAXIMA CORP SAN

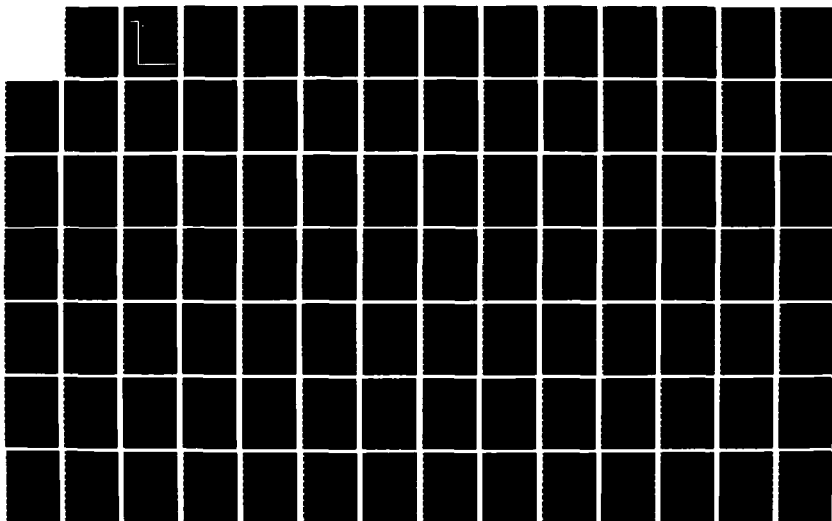
ANTONIO T L DICKINSON ET AL DEC 86 AFHRL-TP-86-32

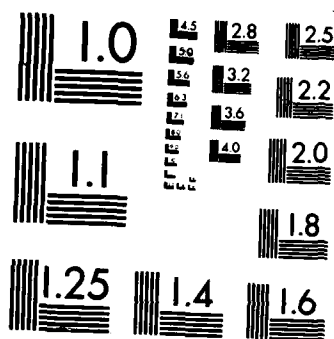
UNCLASSIFIED

F33615-83-C-0030

F/G 5/9

NL





MICROCOPY RESOLUTION TEST CHART  
NATIONAL BUREAU OF STANDARDS-1963-A

12

**AIR FORCE**



**WORK PERFORMANCE RATINGS:  
A META-ANALYSIS OF MULTITRAIT-MULTIMETHOD STUDIES**

Terry L. Dickinson  
Catherine E. Hassett  
Scott I. Tannenbaum

Old Dominion University  
Department of Psychology  
Norfolk, Virginia 23508

TRAINING SYSTEMS DIVISION  
Brooks Air Force Base, Texas 78235-5601

December 1986  
Interim Technical Paper for Period September 1984 - May 1985

Approved for public release; distribution is unlimited.

**LABORATORY**

AD-A174 759

**HUMAN**

**RESOURCES**

DTIC FILE COPY

DTIC  
ELECTE  
DEC 5 1986

**AIR FORCE SYSTEMS COMMAND  
BROOKS AIR FORCE BASE, TEXAS 78235-5601**

B

86 12 04 057

# NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

JERRY HEDGE  
Contract Monitor

HENDRICK W. RUCK, Technical Advisor  
Training Systems Division

GENE A. BERRY, Colonel, USAF  
Chief, Training Systems Division

ADA 104 759

# REPORT DOCUMENTATION PAGE

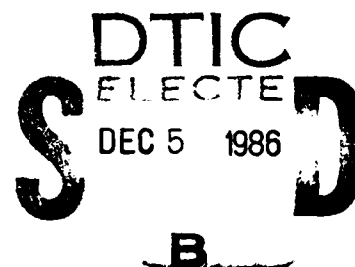
1a. REPORT SECURITY CLASSIFICATION Unclassified			1b. RESTRICTIVE MARKINGS		
2a. SECURITY CLASSIFICATION AUTHORITY			3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.		
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE					
4. PERFORMING ORGANIZATION REPORT NUMBER(S)			5. MONITORING ORGANIZATION REPORT NUMBER(S) AFHRL-TP-86-32		
6a. NAME OF PERFORMING ORGANIZATION The Texas Maxima Corporation		6b. OFFICE SYMBOL (If applicable)		7a. NAME OF MONITORING ORGANIZATION Training Systems Division	
6c. ADDRESS (City, State, and ZIP Code) 8301 Broadway - Suite 212 San Antonio, Texas 78209			7b. ADDRESS (City, State, and ZIP Code) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601		
8a. NAME OF FUNDING / SPONSORING ORGANIZATION Air Force Human Resources Laboratory		8b. OFFICE SYMBOL (If applicable) HQ AFHRL		9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F33615-83-C-0030	
8c. ADDRESS (City, State, and ZIP Code) Brooks Air Force Base, Texas 78235-5601			10. SOURCE OF FUNDING NUMBERS		
PROGRAM ELEMENT NO. 62703F		PROJECT NO. 7734		TASK NO. 08	
				WORK UNIT ACCESSION NO. 24	
11. TITLE (Include Security Classification) Work Performance Ratings: A Meta-Analysis of Multitrait-Multimethod Studies					
12. PERSONAL AUTHOR(S) Dickinson, Terry L.; Hassett, Catherine E.; Tannenbaum, Scott I.					
13a. TYPE OF REPORT Interim		13b. TIME COVERED FROM Sep 84 TO May 85		14. DATE OF REPORT (Year, Month, Day) December 1986	
15. PAGE COUNT 102					
16. SUPPLEMENTARY NOTATION					
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)		
FIELD	GROUP	SUB-GROUP	convergent validity meta-analysis performance ratings		
05	09		discriminant validity method bias		
05	10		job performance multitrait-multimethod		
19. ABSTRACT (Continue on reverse if necessary and identify by block number) An important approach for investigating the quality of performance ratings is an analysis of their multitrait-multimethod (MTMM) properties of convergent validity, method bias, and discriminant validity. The present effort adopted a meta-analytic procedure to provide a quantitative review of the findings of MTMM studies of work performance ratings. Studies were identified for the review by means of a computer-assisted search of the business and social science literature and use of the Social Science Citation Index. A code sheet and code book were developed to specify and define study characteristics that could serve as moderator variables for explaining the differences in MTMM properties. The results indicated that convergent validity was increased through using behavioral dimensions, using example-anchored scales, developing scales rather than modifying existing scales, and involving experts in the development of the rating scales. Method bias was reduced through the use of the same procedures that led to greater convergent validity, as well as by involving raters/ratees in scale development and providing rater training. Finally, discriminant validity was increased through using scales requiring several ratings per dimension, and providing rater training. The review identified several gaps in the literature, as well as deficiencies in the reporting of methods and results. These gaps and deficiencies, together with the quantitative findings, were discussed; and specific areas were suggested for future research.					
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT <input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT <input type="checkbox"/> DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION		
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy A. Perrigo, Chief, STINFO Office			22b. TELEPHONE (Include Area Code) (512) 536-3877		22c. OFFICE SYMBOL AFHRL/TSR

**WORK PERFORMANCE RATINGS:  
A META-ANALYSIS OF MULTITRAIT-MULTIMETHOD STUDIES**

**Terry L. Dickinson  
Catherine E. Hassett  
Scott I. Tannenbaum**

**Old Dominion University  
Department of Psychology  
Norfolk, Virginia 23508**

**TRAINING SYSTEMS DIVISION  
Brooks Air Force Base, Texas 78235-5601**



**Reviewed and submitted by**

**Rodger D. Ballentine, Lt Colonel, USAF  
Chief, Skills Development Branch  
Training Systems Division**

**This publication is primarily a working paper. It is published solely to document work performed.**

## SUMMARY

Important properties that determine the quality of performance ratings are their convergent validity, method bias, and discriminant validity. The present investigation provides a quantitative review of research studies to identify variables that modify the magnitude of these properties. Studies were identified for the review by a computer-assisted search of the literature, and coding procedures were developed to identify moderator variables in each study.

The results indicated that convergent validity was increased through (a) using behavioral dimensions, (b) using example-anchored scales, (c) developing scales rather than revising existing scales, and (d) involving experts in the development of the rating scales. Method bias was reduced through the use of the same procedures that led to greater convergent validity, as well as by involving raters and ratees in scale development, and providing rater training. Discriminant validity was increased through (a) using scales requiring several ratings per dimension, and (b) providing rater training.

Finally, the review also identified gaps in the literature and deficiencies in research methodology. These gaps and deficiencies, together with the quantitative findings, provided recommendations for improving the quality of performance ratings and guiding future research.



Accession	✓
NTIS	
DTIC	
Unannounced	
By	
Dist	
Ann	
Dist	
A-1	

## PREFACE

This work was conducted in partial fulfillment of Contract No. F33615-83-C-0030 awarded to The Texas Maxima Corporation, with the Air Force Human Resources Laboratory (AFHRL). Dr. Jerry W. Hedge served as task monitor. It complements the AFHRL Training Systems Division efforts in job performance criterion development.

# TABLE OF CONTENTS

	<u>Page</u>
I. INTRODUCTION . . . . .	1
Meta-Analysis . . . . .	3
Objectives . . . . .	4
II. METHOD . . . . .	4
Study Domain . . . . .	4
Search Procedures . . . . .	5
Code Sheet . . . . .	7
Coder Training . . . . .	7
Coding Procedure . . . . .	9
Interrater Reliability . . . . .	9
III. RESULTS . . . . .	11
Overview . . . . .	11
Frequencies . . . . .	11
Item and Response Category Reduction . . . . .	12
Subsets of Variables . . . . .	13
Correlations . . . . .	16
Corrections for Study Artifacts . . . . .	26
Analytic Approach . . . . .	29
Regression Analyses . . . . .	31
Subgroup Analyses . . . . .	36

# TABLE OF CONTENTS (concluded)

	<u>Page</u>
IV. DISCUSSION . . . . .	39
Developmental Procedures . . . . .	39
Involvement in Development . . . . .	46
Content and Number of Dimensions . . . . .	46
Rating Format . . . . .	47
Rating Source . . . . .	48
Rater/Ratee Characteristics . . . . .	49
Rating Context: Purpose, Location, and Training . . . . .	49
Prescriptive Recommendations . . . . .	50
Research Questions . . . . .	52
V. CONCLUSIONS . . . . .	53
REFERENCES . . . . .	54
APPENDIX A: ANNOTATED BIBLIOGRAPHY . . . . .	55
APPENDIX B: CODE SHEET WITH STUDY FREQUENCIES FOR RESPONSE CATEGORIES AND MEANS FOR CONTINUOUS ITEMS . . . . .	74
APPENDIX C: CODE BOOK . . . . .	82

# LIST OF TABLES

<u>Table</u>		<u>Page</u>
1	Primary Sources Used to Obtain Studies . . . . .	6
2	Interrogatory Statements Deleted from the Preliminary Code Sheet Due to a Lack of Information in the Multitrait-Multimethod Studies . . . . .	8
3	Subsets of Study Characteristics and the Dependent Variables . . . . .	14
4	Intercorrelations Among Study Characteristics . . . . .	17
5	Correlations Among Dependent Variables . . . . .	26
6	Correlations Between Study Characteristics and Dependent Variables . . . . .	27
7	Average ICC, Observed Variance, Sampling Error Variance, True Variance, and the Percent of Observed Variance Unexplained by Sampling Error for Each of the Dependent Variables . . . . .	32
8	Stepwise Regression by Subset for Convergent Validity . . . . .	34
9	Stepwise Regression by Subset for Method Bias . . . . .	35
10	Stepwise Regression by Subset for Discriminant Validity . . . . .	37
11	Subgroup Analyses for Convergent Validity . . . . .	40
12	Subgroup Analyses for Method Bias . . . . .	42
13	Subgroup Analyses for Discriminant Validity . . . . .	45

## WORK PERFORMANCE RATINGS: A META-ANALYSIS OF MULTITRAIT-MULTIMETHOD STUDIES

### I. INTRODUCTION

Performance evaluation is a universal phenomenon in work environments. Sometimes the performance evaluation is unsystematic and informal in nature, but in many settings a measurement process is institutionalized in the form of a performance rating system. Ratings are the most common procedure for measuring performance, and indeed, for many types of work environments, they are the only practical procedure.

Much of the performance rating research has focused on the psychometric properties of the ratings (e.g., halo, central tendency, leniency) as the indicators of their quality (Landy & Farr, 1980). Another important approach for investigating the quality of performance ratings is an analysis of their multitrait-multimethod (MTMM) properties (Kavanagh, MacKinney, & Wolins, 1971). This approach requires the use of two or more methods for rating two or more traits of work performance. The traits are the performance dimensions on which the ratees are rated. Examples of traits are job requirements (such as planning and organizing) or attributes (such as effort and initiative). The methods can be either the format of the rating scale (e.g., example-anchored and graphic formats) or the source of the ratings (e.g., supervisor, self, and subordinates).

The analysis of MTMM properties focuses on their intercorrelations and provides information concerning several aspects of individual differences in work performance. Convergent validity reflects the extent to which raters agree on the rank-ordering of ratees. The more raters agree, the more likely the ratings describe actual differences between the ratees. Although agreement in the rank-ordering should be due to the amounts of the traits demonstrated by the ratee, it can also be due to the methods of rating. In other words, the presence of convergent validity is not necessarily a positive phenomenon. The nature of the convergence in ratings is important; it should be due primarily to the traits assessed rather than the methods for rating. On the other hand, discriminant validity does reflect the differential ordering of the ratees due to the amounts of the traits demonstrated by the ratees. This is always desirable as work performance is multidimensional,

and ratees should be expected to differ in their rank-ordering from dimension to dimension. Otherwise, one or more dimensions are redundant in describing individual differences in work performance, and these dimensions should be deleted from the rating system.

Method-bias reflects differential ordering of the ratees by the methods used to obtain the ratings. Bias is undesirable if the methods are different rating scale formats. Differential ordering of ratees should be due to individual differences in the amounts of the traits demonstrated by ratees and not due to the format used to make ratings. In contrast, bias may be expected if the methods are sources of ratings. A supervisory source may have observed ratee performance at different times and under different circumstances than a subordinate source; therefore, method bias reflects the sources' differential opportunities to observe performance.

Error reflects residual variation due to sampling and measurement errors. The size of error variation relative to that due to the MTMM properties suggests the extent of the differences between the ratees that cannot be accounted for by the traits and the methods.

In sum, the analysis of MTMM properties provides investigators with information that can be used to assess the degree to which performance ratings reliably order the ratees, differentiate ratees on traits, and are biased by the method of rating.

Lawler (1967) was the first to apply the MTMM approach to the study of performance ratings. The method under consideration was the source of ratings. He investigated the differences between superior ratings, peer ratings, and self-ratings of three traits of managerial performance. Although the superior and peer ratings showed evidence of good convergent and discriminant validity, the self-ratings did not show evidence of either.

Borman and Rosse (1978) investigated the format of the rating instrument. They found no differences in convergent validity across five different rating formats. However, numerical and summated rating formats displayed greater discriminant validity than did the other formats.

Although there have been several reviews of the performance rating literature (e.g., DeCotiis & Petit, 1978; Kavanagh, Borman, Hedge, & Gould, 1986; Landy & Farr, 1980), these reviews have focused on the psychometric properties of ratings. The reviews have offered no prescriptions as to the most effective procedures for enhancing the MTMM properties of work performance ratings. An integration of the MTMM literature would provide information that can be used to identify the variables that influence convergent validity, discriminant validity, and method bias. The purpose of this investigation was to review MTMM studies of work performance ratings and provide recommendations to improve the quality of these ratings.

### Meta-Analysis

The most common form of literature review is the narrative review. However, many problems plague this type of review (see Jackson, 1980). To alleviate some of these problems, the meta-analytic methods of literature review were developed (Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982; Rosenthal, 1978).

A meta-analysis requires a comprehensive literature search to identify and collect all the available studies on the topic of interest. Statistics from the studies are recorded and converted to effect sizes (e.g.,  $r$  and  $d$ ). These effect sizes are analyzed to identify the cumulative results of past research and to draw inferences about the population of potential research results.

There are two different methods used to conduct a meta-analysis. The two methods are similar but are based on different philosophies regarding variation in effect sizes (Mathieu & Tannenbaum, 1983). The Glassian approach (Glass et al., 1981) assumes that the variation in effect sizes is due to moderator variables (i.e., continuous or discrete variables that describe study characteristics). For each study, the statistics and potential moderator variables of interest are recorded. Then, the study statistics are converted to effect sizes and regressed upon the moderator variables to explain the differences between the studies' effect sizes.

The second method of meta-analysis is the Hunter-Schmidt approach (Hunter et al., 1982). As in the Glassian approach, study statistics and potential

moderators are recorded, and the study statistics are converted to effect sizes. However, the Hunter-Schmidt approach is more conservative with regard to moderator variables. It is assumed that some of the variation in the studies' effect sizes is due to methodological artifacts rather than solely due to moderator variables. The Hunter-Schmidt approach corrects for artifacts such as sampling error, unreliability in the measures, and range restriction. After these corrections, the approach examines the variability that remains in the effect sizes. If sufficient unexplained variance remains, then an investigation of moderator variables is considered to be warranted. The correction for artifacts prior to the examination of moderator variables minimizes the likelihood of incorrectly inferring that differences in effect sizes are due to these variables.

### Objectives

The present investigation adopts the meta-analytic procedure developed by Hunter et al. (1982) to provide a quantitative cumulation of the findings of MTMM studies of work performance ratings. The research objectives were: (a) to identify those moderator variables that may affect the quality of performance ratings, (b) to identify gaps in the literature (i.e., variables of interest that have not been examined), (c) to provide recommendations to improve the quality of performance ratings, and (d) to provide direction to guide future research and development (R&D) efforts.

## II. METHOD

### Study Domain

The domain of the meta-analysis was MTMM studies of work performance ratings. A number of MTMM studies were omitted because they did not fit the defined domain. For example, the Borich, Malitz, and Kugle (1978) study did not fit the domain, because it did not examine work performance ratings. Rather, it examined different observation systems. For a similar reason, the Jenkins, Nadler, Lawler, and Cammann (1975) study was omitted. Its rating scales were used to assess job characteristics; therefore, it was not an investigation of work performance.

In addition, studies that did not report the MTMM correlation matrix, presented an inadequate summary of the matrix, or incorrectly calculated the effect sizes (and did not provide sufficient information to allow recalculations) were not included. For example, Wheeler and Knoop (1982) reported a 2 x 2 MTMM matrix when the design seemed to call for a 3 x 3 matrix. An attempt was made to contact authors of such studies in order to obtain the original correlation matrices.

Finally, studies that used raters rather than ratees as the unit of analysis were omitted (e.g., Dickinson & Zellinger, 1980).

### Search Procedures

Multitrait-multimethod studies of work performance ratings were identified by means of a computer-assisted search of the business and social science literature between 1967 and 1985. The year 1967 was chosen to coincide with Lawler's (1967) original application of the MTMM analysis to work performance ratings. The computer search drew on five DIALOG search services: (a) Educational Resources Information Center (ERIC), (b) Management Contents, (c) National Technical Information Service (NTIS), (d) PSYCH INFO, and (e) PSYCH SCAN.

In addition to the computer search, the Social Science Citation Index was used to identify all published studies that cited either Campbell and Fiske (1959) or Kavanagh et al. (1971). Initially, 1,360 individual studies were identified using the Citation Index. However, studies were eliminated that were published in foreign language journals or journals which clearly did not relate to work performance (e.g., Social Psychiatry, Journal of Clinical Psychiatry, Psychosomatic Medicine, Journal of Nervous and Mental Disease). The Citation Index approach identified 506 studies, each of which was examined to determine whether it would fit the defined domain of work performance ratings for the meta-analysis.

Several published reviews of the performance rating literature were also examined. Only three additional studies were identified by these reviews. Table 1 represents a list of the primary reference sources.

In order to obtain possible unpublished studies for inclusion in the meta-analysis, a letter was written to

Table 1. Primary Sources Used to Obtain Studies

Source	Year covered
1. <u>DIALOG Search Services</u>	
ERIC	1967--1985
Management Contents	1974--1985
NTIS	1967--1985
PSYCH INFO	1967--1985
PSYCH SCAN	1967--1985
2. <u>Social Science Citation Index</u>	
Campbell and Fiske (1959)	1967--1985
Kavanagh, MacKinney, and Wolins (1971)	1971--1985
3. <u>Literature Reviews</u>	
DeCotiis and Petit (1978)	
Jacobs, Kafry, and Zedeck (1979)	
Kane and Lawler (1978)	
Kavanagh, Borman, Hedge, and Gould (1986)	
Kingstrom and Bass (1981)	
Landy and Farr (1980)	
Saal, Downey, and Lahey (1980)	
Schwab, Heneman, and DeCotiis (1975)	

all authors of identified published studies. This letter asked each author whether he or she possessed an unpublished MTMM study or knew of another researcher who might possess such a manuscript. No usable studies were identified with this procedure. Thus, every effort was

made to cover the literature domain. All articles included in the meta-analysis are presented in an annotated bibliography (see Appendix A).

### Code Sheet

The literature reviews reported in Table 1 served as the basis for an identification of study characteristics to be included in a code sheet. Although these reviews focused on the psychometric properties of performance ratings, they provided a directive function for the meta-analysis.

Five broad topic areas were used initially in developing the code sheet. Study characteristics were classified according to: (a) the nature of the traits rated, (b) the methods used for rating, (c) the nature of the ratees, (d) the nature of the raters, and (e) the context of the ratings.

The preliminary code sheet consisted of interrogatory statements for each of the five topic areas. Next, the MTMM studies were reviewed to assess the availability of information regarding the study characteristics. Several study characteristics were eliminated due to a lack of information. These characteristics dealt with rater and ratee age, race, and socioeconomic status, as well as raters' knowledge and acceptance of the performance rating system and rater-ratee interactions. See Table 2 for a list of the deleted interrogatory statements.

Items were written for the code sheet using each of the remaining statements as stems. All items adhered to one of two formats: (a) a check was to be placed beside the appropriate response category, or (b) a number was to be provided.

Finally, during the coder training process, the code sheet was reviewed and revised to ensure that all items and their response categories were clearly stated and understandable. The final code sheet is presented in Appendix B.

### Coder Training

As the code sheet was being revised, the coders discussed the precise definitions of the items and of each response category. The definitions were recorded

Table 2. Interrogatory Statements Deleted from the Preliminary Code Sheet Due to a Lack of Information in the Multitrait-Multimethod Studies

---

How many points comprised each scale?  
What was the socioeconomic status of the ratees?  
What was the age of the ratees?  
What was the race of the ratees?  
What was the socioeconomic status of the raters?  
What was the age of the raters?  
What was the race of the raters?  
What was the education level of the raters?  
To what degree was there rater/ratee sex congruence?  
To what degree was there rater/ratee race congruence?  
How many hours of rater training did the raters receive?  
Did "booster" training occur?  
What was the knowledge or understanding of the job by the raters?  
What was the raters' acceptance of the performance rating system?  
What was the level of interpersonal trust between raters and ratees?  
What was the level of rater/ratee conflict?  
What was the level of rater/ratee acquaintanceship or friendship?  
What was the degree to which raters had opportunity to observe job performance?

---

in a code book to be referred to during the coding process.

The purpose of the code book was to guide coders and to increase interrater reliability. The code book answered questions regarding differences between response categories, location of responses, and correct calculations for items requiring a numerical response. The code book also identified items where more than one response could be appropriate.

Several studies were coded using a preliminary code book. Ambiguities, questions, and clarifications were noted and discussed by all coders. The code book was revised to reflect these discussions. The final code book is presented in Appendix C.

### Coding Procedure

Using the procedures established in the code book, three coders independently coded all items for each of the 31 correlation matrices. The coders then met and discussed the completed code sheets. All discrepancies were resolved by a consensus decision.

A consensus decision process was used rather than a statistical pooling of responses for several reasons. First, since most of the coding was discrete in nature, scores resulting from pooling would have been meaningless. Second, the consensus process served as continual training for coders. Third, in most instances, there were specific statements from the study that could be quoted to support a coding response. When one coder noted a particular statement from a study, the other coders either agreed with the coder or contested the coding by quoting another statement from the study. The ability of a minority opinion coder to quote a particular statement increased the thoroughness of the coding task by reducing errors of omission and rendered the consensus decision process more desirable than statistical pooling.

For each item and its response categories, four pieces of information were recorded: the independent responses made by the three coders and the group consensus decision. The three independent codings were used for the reliability analysis; the consensus decisions were used for the remaining analyses.

### Interrater Reliability

A number of interrater reliability indices have been suggested in the literature (e.g., Cohen's kappa, average interrater correlation, percent agreement, intraclass correlations, and Spearman-Brown formula). The applicability of the indices varies for different purposes and types of data. For meta-analytic purposes, a reliability index should describe coder agreement and not coder consistency. With the latter type of index, coders could provide different responses and still demonstrate perfect reliability. The average interrater correlation and the Spearman-Brown formula do not directly assess agreement among coders (Jones, Johnson, Butler, & Main, 1983); therefore, they were not applicable for this study.

Two types of responses were required in coding: dichotomous for discrete items (e.g., was a factor analysis conducted?) and numerical for continuous items (e.g., what was the number of ratees per rater?). The distributions of the dichotomous responses were often extremely skewed and, thus, contained little variance. This is a common occurrence with dichotomous responses, particularly in meta-analyses (cf. Bullock & Svyantek, 1985). Cohen's kappa (1968) and percent agreement are two indices of agreement that may be appropriate for use with dichotomous responses. However, in an empirical investigation of the most common indices of agreement, Jones et al. (1983) found that "under conditions of restricted variance in the ratings, only percent of agreement appeared to provide an accurate indication of rating overlap" (p. 515). Burton (1981) also found that "none of the complex coefficients of agreement are appropriate when there is little variation in the data" (p. 956), and she found Cohen's kappa to be overly sensitive to skewed data. For the above reasons, average percent of pairwise agreement was computed for the three coders for each of the dichotomous response categories.

A different analytic approach was used to determine coder agreement for the continuous items. Two intraclass correlation coefficients (ICCs) were computed for each item: ICC (2,1) and ICC (2,k), where k is the number of coders (Shrout & Fleiss, 1979). These particular formulas were chosen because they reflect coder agreement and not simply coder consistency. ICC (2,1) is the reliability for a single, typical coder; ICC (2,3) is the reliability for the three coders.

The reliability statistics indicated that a high level of interrater reliability was obtained with the coding procedure. The percentages of agreement for the dichotomous response categories ranged from 70% to 94.5%. After deletion of those response categories that exhibited no variability (i.e., the consensus process indicated that for all 31 matrices the group response was "1" or "0"), the mean average percent agreement was 92.8%. The ICCs for the five continuous items ranged from .404 to .982 for ICC (2,1) and from .670 to .993 for ICC (2,3). Average values were .741 for ICC (2,1) and .877 for ICC (2,3).

### III. RESULTS

#### Overview

Studies were coded according to 6 continuous items and 114 response categories. These study characteristics were reduced to 4 continuous items and 33 response categories on the basis of the obtained data. This reduction resulted in eight subsets of moderator variables that could be analyzed to explain variation in the three effect sizes of convergent validity, method bias, and discriminant validity.

A three-step analytic procedure was used. First, the effect sizes were computed as ICCs. On the average, the ratings exhibited fairly high convergent validity (.356), moderate method bias (.223), and relatively low discriminant validity (.128). The ICCs were corrected only for the artifact of sampling error; as discussed later, corrections for the artifacts of restriction of range and unreliability were considered to be unnecessary. Sufficient variance remained in each set of ICCs after the sampling error correction to warrant searching for moderator variables. Second, a stepwise regression analysis was used to identify potential moderator variables for the three effect sizes (Hull & Nie, 1981). For convergent validity, 14 study characteristics were identified as potential moderators; for method bias, 17; and for discriminant validity, 7. Finally, for each study characteristic that entered a regression analysis, weighted means and variances were evaluated. This was done to assure significant reduction in true variance in the subgroup of studies that possessed the characteristic compared to that of the total group. Only 8 potential moderators were eliminated by the subgroup analysis, leaving 30 moderator relationships with the ICCs of convergent validity, method bias, and discriminant validity.

#### Frequencies

Frequency data provided information about the study domain (see Appendix B for the frequencies of each response category). For example, most MTMM studies were conducted in industrial or classroom settings. In addition, frequency data helped to identify gaps in the previous research. Two gaps were a lack of information regarding purpose of rating and rater/ratee characteristics. Only 7 studies directly reported the

purpose for the ratings. Nineteen studies did not provide information about rater sex, and 21 studies did not provide information about ratee sex. All other rater/ratee characteristics (e.g., age) had to be eliminated prior to the completion of the code sheet. In addition, only three studies reported the use of rater training. For this reason, it was impossible to examine which features of rater training affected MTMM properties.

As noted in the method section, other variables that may have an effect on rating quality (e.g., rater's opportunity to observe job performance) did not appear as items on the current code sheet. These variables, reported in Table 2, demonstrated insufficient variability and, thus, were not investigated in the current study. However, this does not indicate that the excluded variables are unimportant. For example, a recent meta-analysis of psychometric properties specifically examined ratee race effects and found that raters gave higher ratings to ratees of their own race (Kraiger & Ford, 1985). Clearly, it is necessary to consider the variables that MTMM research studies failed to examine or describe in order to understand the research domain and to identify needs for future research.

#### Item and Response Category Reduction

The code sheet was composed of 6 continuous items and 114 response categories to describe study characteristics. They were reduced to 4 continuous items and 33 response categories on the basis of frequencies, intercorrelations, and conceptual similarity. These 37 study characteristics were used for subsequent analyses. The reduction of the items and response categories is described below.

Any response category that had a zero frequency was disregarded. Furthermore, response categories with few occurrences were merged to allow for meaningful interpretations. Specifically, those studies that reported the use of either job descriptions, surveys, discussions with subject-matter experts, or retranslation (see Smith & Kendall, 1963) were reclassified as "using a systematic process to identify performance dimensions." Studies that reported the use of either behaviorally anchored rating scales (BARS) or behavioral expectation scales (BES) were merged together as "example-anchored behavior scales"; the two studies that reported the use of mixed standard scales were

reclassified and grouped with those studies that used "other" types of scales. For the item concerning purpose of rating, the response categories of criterion-related validity, basic research, and employee growth and development were collapsed into one category called "ratings for non-administrative purposes." Furthermore, all those studies that reported that ratees were involved in the development of the rating system reported that the raters were also involved; thus, these two categories were merged. With regard to the content of the performance dimensions, all studies were reclassified into one of three categories: strictly behavioral dimensions, a mix of behavioral and trait dimensions, or strictly trait or "other" dimensions.

Items 13a and 14a of the code sheet were indicators of research design quality. They addressed the degree to which rater variance and ratee variance were confounded in the individual ratings of ratee performance (see items 13a and 14a in Appendix C for a more detailed description). All studies were reclassified on the basis of joint frequencies between the response categories of items 13a and 14a as either being completely mixed designs (greatest confounding of variance) or having some nested or crossed procedures (less confounding of variance).

Lastly, insufficient reporting of rater and ratee gender necessitated the following dichotomies: all male and all female raters versus raters whose sex was not explicitly stated, all male raters versus all female raters and those raters whose sex was not explicitly stated, all male ratees versus all female ratees and those ratees whose sex was not explicitly stated, and all female ratees versus all male ratees and those ratees whose sex was not explicitly stated. Moreover, most studies had to be categorized as "not stated" (N = 19 for rater sex; N = 21 for ratee sex). Thus, the effects of having all female raters or a mixed group of ratees could not be examined due to insufficient variability.

#### Subsets of Variables

Nine meaningful subsets of variables were examined: eight subsets of study characteristics (i.e., items and response categories) and one set of dependent variables. The subsets are listed in Table 3. The dependent variables were the effect sizes of convergent validity,

Table 3. Subsets of Study Characteristics and the  
Dependent Variables

---

1. Procedures to Develop Dimensions
  - A. Systematic development of performance dimensions
  - B. Derivation from existing scale
  - C. Factor analysis
  - D. Expert prescription
2. Involvement in the Development of the Rating Scale
  - A. Raters or ratees involved in development
  - B. Experts involved in development
  - C. Existing scale used (no involvement)
  - D. Existing scale modified (some involvement possible)
3. Content and Number of Dimensions
  - A. Dimensions strictly behavioral
  - B. Some dimensions behavioral, some trait
  - C. Trait dimensions and/or "other"
  - D. Specific content (1) vs. general (0)
  - E. Number of dimensions
  - F. Number of ratings per dimension
4. Rating Format
  - A. BARS/BES
  - B. Graphic
  - C. MSS/"other"
5. Rating Source
  - A. 1st level supervisors
  - B. 2nd level supervisors
  - C. Peers
  - D. Self
  - E. Subordinates
  - F. Students

Table 3. (concluded)

---

6. Rater-Ratee Characteristics: Gender and Ratio

- A. Both male and female raters
- B. All male raters
- C. All male ratees
- D. All female ratees
- E. Number of ratees per rater

7. Rating Context: Purpose, Location, and Rater Training

- A. Ratings for non-administrative purposes
- B. Private industry
- C. Military
- D. Academia
- E. Public sector
- F. Rater training

8. Study Design

- A. "Method" in MTMM matrix: source of rating (1) vs. format (0)
- B. Completely mixed design (1) vs. some nesting and/or crossing (0)
- C. Number of ratees

9. Dependent Variables

- A. Convergent validity
- B. Method bias
- C. Discriminant validity

---

Note. The continuous variables were 3E, 3F, 6E, 8C, 9A, 9B, and 9C.

method bias, and discriminant validity. The effect sizes were computed as ICCs from an analysis of variance (ANOVA) of the MTMM correlation matrices.

Most authors of MTMM studies have followed the suggestion of Kavanagh et al. (1971) to calculate ICCs as the ratio of a source's variance component to that component plus the error variance component. However, as first noted by Bartko (1966), an ICC should be defined as the ratio of a source's variance component to

the sum of all variance components. Using the sum of all components as the denominator expresses an ICC as a proportion of the total variation accounted for in the study. In the present research, Bartko's definition was used to calculate ICCs.

For studies that provided the MTMM correlation matrix, the ANOVA and the ICCs were computed using a computer program. For the 10 studies that did not report the correlation matrix and included only an ANOVA summary table, the ICCs were computed on the basis of the mean squares reported in that table.

### Correlations

Correlations were computed among the study characteristics, among the dependent variables, and between the study characteristics and the dependent variables. Table 4 presents the intercorrelations among the 37 study characteristics.

As shown in Table 4, there was collinearity among the study characteristics. Some of the high correlations were expected. For example, studies conducted in an academic setting were likely to involve student raters ( $r = .79$ ). Other high correlations were not anticipated. For example, modifications of existing scales were usually developed via factor analysis ( $r = .82$ ). As discussed below, it was necessary to consider the intercorrelations among the study characteristics to be able to interpret the results of the meta-analysis.

Table 5 displays the intercorrelations among the dependent variables. All three dependent variables were negatively correlated, with the strongest negative correlation between discriminant validity and method bias ( $r = -.56$ ).

Of course, negative correlations are to be expected among the ICCs. Increases in the proportion of total variance explained by one dependent variable decrease the magnitude of the proportion of total variance explained by other variables. This fact suggests that study characteristics which correlate with one dependent variable are likely to correlate with another dependent variable in the opposite direction.

Table 4. Intercorrelations Among Study Characteristics

Study variable	1A	1B	1C	1D
(1A) systematic development	1.00			
(1B) derivation from ex. scale	-.54*	1.00		
(1C) factor analysis	-.49*	.73*	1.00	
(1D) expert prescription	.31*	-.14	-.03	1.00
(2A) raters/ratees involved	.38*	-.19	-.07	.45*
(2B) experts involved	.05	-.02	-.07	.67*
(2C) existing scale used	-.16	.52*	.02	-.17
(2D) existing scale modified	-.49*	.73*	.82*	-.03
(3A) behavioral dimensions	.72*	-.24	-.52*	.17
(3B) behav/trait mix	-.49*	.08	.26	-.24
(3C) traits and "other" dims	-.40*	.22	.39*	.05
(3D) spec content vs. general	.60*	.01	-.01	.08
(3E) number of dimensions	.41*	-.05	.06	.14
(3F) number of ratings per dim	-.27	.52*	.24	-.15
(4A) BARS/BES	.65*	-.44*	-.32*	.14
(4B) graphic	-.75*	.63*	.56*	-.07
(4C) MSS/"other"	.50*	-.13	-.14	.34*
(5A) 1st level supervisor	.28	-.49*	-.32*	-.01
(5B) 2nd level supervisor	.48*	-.31*	-.27	-.17
(5C) peers	-.21	-.22	-.04	.07
(5D) self	-.49*	.21	-.04	-.12
(5E) subordinate	-.35*	.52*	.71*	.09
(5F) students	-.35*	.52*	.25	-.17
(6A) male and female raters	-.30	.21	-.18	-.14
(6B) all male raters	.28	-.33*	-.18	-.28
(6C) all male ratees	-.31*	-.02	.08	-.15
(6D) all female ratees	.53*	-.38*	-.28	-.22
(6E) number ratees per rater	.47*	-.78*	-.63*	-.48*
(7A) nonadministrative purpose	-.28	-.19	-.26	.01
(7B) private industry	.17	-.28	.03	.05
(7C) military	.03	.08	.17	.24
(7D) academia	-.28	.32*	.13	-.21
(7E) public sector	.05	-.02	-.07	.01
(7F) rater training	.36*	-.01	-.18	.15
(8A) method: source vs. format	-.36*	.24	.18	-.75*
(8B) mixed vs. nested/crossed	.69*	-.41*	-.29	.05
(8C) number of ratees	.09	.00	.20	.38*

Table 4. (continued)

Study variable	2A	2B	2C	2D
(1A)systematic development				
(1B)derivation from ex. scale				
(1C)factor analysis				
(1D)expert prescription				
(2A)raters/ratees involved	1.00			
(2B)experts involved	.38*	1.00		
(2C)existing scale used	-.19	-.19	1.00	
(2D)existing scale modified	-.07	.13	-.21	1.00
(3A)behavioral dimensions	.22	.05	.11	-.36*
(3B)behav/trait mix	-.26	-.26	-.21	.26
(3C)traits and "other" dims	.01	.23	.09	.18
(3D)spec content vs. general	.12	-.26	.02	-.01
(3E)number of dimensions	-.11	.00	-.27	.16
(3F)number of ratings per dim	-.27	-.23	.54*	.15
(4A)BARS/BES	.27	-.10	-.23	-.32*
(4B)graphic	-.31*	.02	.21	.56*
(4C)MSS/"other"	.08	.08	.21	-.32*
(5A)1st level supervisor	.03	.03	-.30	-.32*
(5B)2nd level supervisor	-.05	-.22	.09	-.43*
(5C)peers	.36*	.01	-.06	-.21
(5D)self	-.16	.19	.15	.12
(5E)subordinate	.06	.06	-.15	.71*
(5F)students	-.19	-.19	.43*	.25
(6A)male and female raters	-.16	.12	.53*	-.18
(6B)all male raters	-.31*	-.31*	-.03	-.35*
(6C)all male ratees	-.17	-.15	.20	-.18
(6D)all female ratees	-.25	-.22	-.20	-.28
(6E)number ratees per rater	-.17	-.52*	N/A	-.78*
(7A)nonadministrative purpose	-.24	.38*	-.19	-.07
(7B)private industry	-.05	-.21	-.42*	.02
(7C)military	.20	.20	.29	-.14
(7D)academia	-.03	-.24	.30	.13
(7E)public sector	-.03	.17	.30	-.26
(7F)rater training	.12	.12	.20	-.18
(8A)method: source vs. format	-.12	-.39*	.13	.18
(8B)mixed vs. nested/crossed	.12	-.21	-.04	-.44*
(8C)number of ratees	.22	.28	-.20	.16

Table 4. (continued)

Study variable	3A	3B	3C	3D
(1A)systematic development				
(1B)derivation from ex. scale				
(1C)factor analysis				
(1D)expert prescription				
(2A)raters/ratees involved				
(2B)experts involved				
(2C)existing scale used				
(2D)existing scale modified				
(3A)behavioral dimensions	1.00			
(3B)behav/trait mix	-.68*	1.00		
(3C)traits and "other" dims	-.55*	-.24	1.00	
(3D)spec content vs. general	.53*	.13	-.53*	1.00
(3E)number of dimensions	.23	.06	-.37*	.47*
(3F)number of ratings per dim.	.04	.15	-.22	.42*
(4A)BARS/BES	.47*	-.32*	-.26	.38*
(4B)graphic	-.42*	.25	.28	-.44*
(4C)MSS/"other"	.32*	-.32*	-.06	.21
(5A)1st level supervisor	-.05	-.13	.21	-.32*
(5B)2nd level supervisor	.36*	-.43*	.01	.05
(5C)peers	-.44*	.29	.26	-.48*
(5D)self	-.30*	.29	.07	-.36*
(5E)subordinate	-.29	.02	.35*	.02
(5F)students	-.09	.25	-.17	.25
(6A)male and female raters	.04	-.18	.15	-.04
(6B)all male raters	.07	-.01	-.09	.17
(6C)all male ratees	-.44*	.34*	.20	-.33*
(6D)all female ratees	.38*	-.28	-.20	.30
(6E)number ratees per rater	.47*	N/A	-.47*	.14
(7A)nonadministrative purpose	-.28	.13	.23	-.64*
(7B)private industry	-.06	.18	-.13	.18
(7C)military	-.06	-.14	.24	-.14
(7D)academia	-.11	.13	.01	.12
(7E)public sector	.05	-.26	.23	-.26
(7F)rater training	.26	-.18	-.14	.21
(8A)method: source vs. format	-.26	.18	.14	-.21
(8B)mixed vs. nested/crossed	.34*	-.13	.31*	.48*
(8C)number of ratees	-.10	.14	-.02	.18

Table 4. (continued)

Study variable	3E	3F	4A	4B
(1A)systematic development				
(1B)derivation from ex. scale				
(1C)factor analysis				
(1D)expert prescription				
(2A)raters/ratees involved				
(2B)experts involved				
(2C)existing scale used				
(2D)existing scale modified				
(3A)behavioral dimensions				
(3B)behav/trait mix				
(3C)traits and "other" dims				
(3D)spec content vs. general				
(3E)number of dimensions	1.00			
(3F)number of ratings per dim	-.15	1.00		
(4A)BARS/BES	.39*	-.49*	1.00	
(4B)graphic	-.18	.32	-.57*	1.00
(4C)MSS/"other"	.07	.19	-.18	-.28
(5A)1st level supervisor	.11	-.82*	.10	-.34*
(5B)2nd level supervisor	.14	-.32	.44*	-.37*
(5C)peers	-.47*	-.39*	-.25	.02
(5D)self	-.23	.28	-.41*	.44*
(5E)subordinate	.16	-.17	-.23	.40*
(5F)students	-.09	.82*	-.23	.40*
(6A)male and female raters	-.22	.65*	-.19	.12
(6B)all male raters	.20	-.18	.27	-.48*
(6C)all male ratees	-.38*	-.01	-.20	-.11
(6D)all female ratees	.49*	-.18	.45*	-.50*
(6E)number ratees per rater	.22	.35	.29	-.47*
(7A)nonadministrative purpose	-.17	-.29	-.29	.34*
(7B)private industry	.40*	-.46*	.39*	-.29
(7C)military	-.12	-.13	-.15	.01
(7D)academia	-.19	.55*	-.10	.34*
(7E)public sector	-.35*	.16	-.29	.02
(7F)rater training	-.04	-.05	.06	-.10
(8A)method: source vs. format	-.22	-.11	-.31*	.10
(8B)mixed vs. nested/crossed	.21	-.07	.54*	-.68*
(8C)number of ratees	.22	.25	.19	-.07

Table 4. (continued)

Study variable	4C	5A	5B	5C
(1A)systematic development				
(1B)derivation from ex. scale				
(1C)factor analysis				
(1D)expert prescription				
(2A)raters/ratees involved				
(2B)experts involved				
(2C)existing scale used				
(2D)existing scale modified				
(3A)behavioral dimensions				
(3B)behav/trait mix				
(3C)traits and "other" dims				
(3D)spec content vs. general				
(3E)number of dimensions				
(3F)number of ratings per dim				
(4A)BARS/BES				
(4B)graphic				
(4C)MSS/"other"	1.00			
(5A)1st level supervisor	.29	1.00		
(5B)2nd level supervisor	.14	.22	1.00	
(5C)peers	.07	.34*	-.27	1.00
(5D)self	-.25	-.36*	-.41*	.11
(5E)subordinate	-.23	-.06	-.31*	-.06
(5F)students	-.23	-.79*	-.31*	-.27
(6A)male and female raters	-.19	-.39*	-.26	-.23
(6B)all male raters	.11	.31	.51*	.01
(6C)all male ratees	.05	.17	-.04	.47*
(6D)all female ratees	.08	.25	.61*	-.35*
(6E)number ratees per rater	.00	.37	.93*	-.34
(7A)nonadministrative purpose	-.10	.24	-.22	.19
(7B)private industry	-.21	.21	.19	-.07
(7C)military	.45*	.13	.06	.38*
(7D)academia	-.29	-.59*	-.22	.01
(7E)public sector	.46*	.24	.11	.01
(7F)rater training	.31*	.16	.19	.23
(8A)method: source vs. format	-.31*	.12	.04	-.23
(8B)mixed vs. nested/crossed	.39*	.21	.46*	-.07
(8C)number of ratees	-.01	-.38*	-.01	-.21

Table 4. (continued)

Study variable	5D	5E	5F	6A
(1A)systematic development				
(1B)derivation from ex. scale				
(1C)factor analysis				
(1D)expert prescription				
(2A)raters/ratees involved				
(2B)experts involved				
(2C)existing scale used				
(2D)existing scale modified				
(3A)behavioral dimensions				
(3B)behav/trait mix				
(3C)traits and "other" dims				
(3D)spec content vs. general				
(3E)number of dimensions				
(3F)number of ratings per dim				
(4A)BARS/BES				
(4B)graphic				
(4C)MSS/"other"				
(5A)1st level supervisor				
(5B)2nd level supervisor				
(5C)peers				
(5D)self	1.00			
(5E)subordinate	-.27	1.00		
(5F)students	.56*	-.15	1.00	
(6A)male and female raters	.47*	-.13	.53*	1.00
(6B)all male raters	-.44*	-.25	-.25	-.21
(6C)all male ratees	-.22	-.13	-.13	-.09
(6D)all female ratees	-.33*	-.20	-.20	-.13
(6E)number ratees per rater	-.34	-.63*	N/A	N/A
(7A)nonadministrative purpose	.54*	-.19	-.19	.11
(7B)private industry	-.48*	.35*	-.42*	-.36*
(7C)military	.10	-.10	-.10	-.09
(7D)academia	.54*	.19	.79*	.39*
(7E)public sector	-.16	-.19	-.19	.12
(7F)rater training	-.23	-.13	-.13	.11
(8A)method: source vs. format	.23	.13	.13	.11
(8B)mixed vs. nested/crossed	-.62*	-.42*	-.23	-.36*
(8C)number of ratees	-.11	.06	.03	-.12

Table 4. (continued)

Study variable	6B	6C	6D	6E
(1A)systematic development				
(1B)derivation from ex. scale				
(1C)factor analysis				
(1D)expert prescription				
(2A)raters/ratees involved				
(2B)experts involved				
(2C)existing scale used				
(2D)existing scale modified				
(3A)behavioral dimensions				
(3B)behav/trait mix				
(3C)traits and "other" dims				
(3D)spec content vs. general				
(3E)number of dimensions				
(3F)number of ratings per dim				
(4A)BARS/BES				
(4B)graphic				
(4C)MSS/"other"				
(5A)1st level supervisor				
(5B)2nd level supervisor				
(5C)peers				
(5D)self				
(5E)subordinate				
(5F)students				
(6A)male and female raters				
(6B)all male raters	1.00			
(6C)all male ratees	-.51*	1.00		
(6D)all female ratees	.76*	-.17	1.00	
(6E)number ratees per rater	.56*	N/A	.56*	1.00
(7A)nonadministrative purpose	-.31*	-.15	-.22	-.37
(7B)private industry	.44*	.07	.44*	-.03
(7C)military	.12	.36*	-.13	-.10
(7D)academia	-.31	-.17	-.25	N/A
(7E)public sector	-.13	.15	-.22	.51*
(7F)rater training	-.21	-.11	-.17	.20
(8A)method: source vs. format	.21	.11	.17	.31
(8B)mixed vs. nested/crossed	.58*	.29	.44*	.50*
(8C)number of ratees	-.23	-.05	-.21	-.50*

Table 4. (continued)

Study variable	7A	7B	7C	7D
(1A)systematic development				
(1B)derivation from ex. scale				
(1C)factor analysis				
(1D)expert prescription				
(2A)raters/ratees involved				
(2B)experts involved				
(2C)existing scale used				
(2D)existing scale modified				
(3A)behavioral dimensions				
(3B)behav/trait mix				
(3C)traits and "other" dims				
(3D)spec content vs. general				
(3E)number of dimensions				
(3F)number of ratings per dim				
(4A)BARS/BES				
(4B)graphic				
(4C)MSS/"other"				
(5A)1st level supervisor				
(5B)2nd level supervisor				
(5C)peers				
(5D)self				
(5E)subordinate				
(5F)students				
(6A)male and female raters				
(6B)all male raters				
(6C)all male ratees				
(6D)all female ratees				
(6E)number ratees per rater				
(7A)nonadministrative purpose	1.00			
(7B)private industry	-.21	1.00		
(7C)military	-.13	-.29	1.00	
(7D)academia	-.03	-.54*	-.13	1.00
(7E)public sector	.17	-.54*	.20	-.24
(7F)rater training	.12	-.08	.09	.16
(8A)method: source vs. format	-.12	-.08	.09	.16
(8B)mixed vs. nested/crossed	-.38*	.22	-.03	-.21
(8C)number of ratees	-.05	.14	-.10	-.09

Table 4. (concluded)

Study variable	7 E	7 F	8 A	8 B
(1A)systematic development				
(1B)derivation from ex. scale				
(1C)factor analysis				
(1D)expert prescription				
(2A)raters/ratees involved				
(2B)experts involved				
(2C)existing scale used				
(2D)existing scale modified				
(3A)behavioral dimensions				
(3B)behav/trait mix				
(3C)traits and "other" dims				
(3D)spec content vs. general				
(3E)number of dimensions				
(3F)number of ratings per dim				
(4A)BARS/BES				
(4B)graphic				
(4C)MSS/"other"				
(5A)1st level supervisor				
(5B)2nd level supervisor				
(5C)peers				
(5D)self				
(5E)subordinate				
(5F)students				
(6A)male and female raters				
(6B)all male raters				
(6C)all male ratees				
(6D)all female ratees				
(6E)number ratees per rater				
(7A)nonadministrative purpose				
(7B)private industry				
(7C)military				
(7D)academia				
(7E)public sector	1.00			
(7F)rater training	.39*	1.00		
(8A)method: source vs. format	-.12	-.26	1.00	
(8B)mixed vs. nested/crossed	.12	.30	-.30	1.00
(8C)number of ratees	.03	.00	-.61*	.25

Note. N/A = insufficient data to compute the correlation.

\* $p < .05$ .

**Table 5. Correlations Among Dependent Variables**

	Convergent validity	Method bias	Discriminant validity
Convergent validity	1.00		
Method bias	-.35*	1.00	
Discriminant validity	-.16	-.56*	1.00

\* $p < .05$ .

Table 6 shows the correlations between the study characteristics and the dependent variables. These correlations suggest that moderator variables may explain variation in the effect sizes. However, the apparent variation may be due to study artifacts, and so, corrections for artifacts were considered.

#### Corrections for Study Artifacts

In their approach to meta-analysis, Hunter et al. (1982) recommended that, whenever possible, corrections should be made for sampling error, as well as for unreliability and range restriction. Of the three corrections, sampling error usually accounts for the majority of the spurious variance (Schmitt, Gooding, Noe, & Kirsch, 1984).

In the current study, only the correction for sampling error was necessary. For MTMM studies of performance ratings, information about range restriction is not meaningful. The entire population of employees or a large unrestricted sample of this population was rated.

In their meta-analysis of selection research, Hunter et al. (1982) corrected for unreliability in the work performance measure (e.g., performance ratings, work samples), but they did not correct for

**Table 6. Correlations Between Study Characteristics and Dependent Variables**

Study characteristics	<u>Dependent variables</u>		
	CV	MB	DV
<b>1. <u>Procedures to Develop Dimensions</u></b>			
A. Systematic development	.44*	-.41*	-.10
B. Derivation from existing scale	-.37*	-.17	.40*
C. Factor analysis	-.29	-.03	.33*
D. Expert prescription	.34*	-.39*	.17
<b>2. <u>Involvement in Development of Rating Scale</u></b>			
A. Raters/ratees involved	-.12	-.21	-.01
B. Experts involved	.11	-.12	.06
C. Existing scale used	.06	-.26	.24
D. Existing scale modified	-.47*	.02	.26
<b>3. <u>Content and Number of Dimensions</u></b>			
A. Strictly behavioral	.30*	-.52*	-.09
B. Some behavioral, some trait	-.44*	.46*	.17
C. Trait and/or "other"	.10	.16	-.07
D. Specific content vs. general	.08	-.34*	.25
E. Number of dimensions	.17	-.24	-.05
F. Number of ratings per dimension	-.32	-.30	.63*
<b>4. <u>Rating Format</u></b>			
A. BARS/BES	.52*	-.40*	-.09
B. Graphic	-.31*	-.06	.35*
C. MSS/"other"	.17	-.22	.03
<b>5. <u>Rating Source</u></b>			
A. 1st level supervisors	.31*	.27	-.61*
B. 2nd level supervisors	.51*	-.28	-.20
C. Peers	-.23	.32*	-.16
D. Self	-.56*	.18	.29
E. Subordinates	-.15	.14	-.14
F. Students	-.42*	-.28	.68*

Table 6. (concluded)

Study characteristics	<u>Dependent variables</u>		
	CV	MB	DV
<b>6. <u>Rater-Ratee Characteristics: Gender and Ratio</u></b>			
A. Both male and female raters	-.09	.03	.12
B. All male raters	.36*	-.01	-.21
C. All male ratees	.12	.32*	-.01
D. All female ratees	.33*	-.19	-.26
E. Number of ratees per rater	.42	-.05	-.44
<b>7. <u>Rating Context: Purpose, Location, and Rater Training</u></b>			
A. Non-administrative purposes	-.06	.28	-.01
B. Private industry	.22	.13	-.33*
C. Military	.01	-.05	.01
D. Academia	-.37*	-.24	.56*
E. Public Sector	.27	.05	-.08
F. Rater training	.28	-.26	.26
<b>8. <u>Study Design</u></b>			
A. Method: source vs. format	-.54*	.39*	-.27
B. Mixed vs. nested/crossed	.45*	-.32*	.10
C. Number of ratees	.10	-.08	.27

Note. CV = convergent validity; MB = method bias; and DV = discriminant validity. All correlations were based on a sample size of 31 except the following categories: 3D where N = 29, 3F where N = 21, 6C and 6D where N = 30, and 6E where N = 14.

\* $p < .05$ .

unreliability in the predictors (e.g., selection tests). They argued that for validation purposes, predictors should not be corrected, because they are used in imperfect form for hiring. On the other hand, the work performance measure must be corrected, because the degree to which an organization benefits from valid predictors is indicated by actual performance and not by the unreliable work performance measure. Therefore, the appropriate validity coefficient is a coefficient which is corrected for unreliability in the work performance measure but not in the predictors.

Their argument is cogent, and a similar rationale could be applied to the current research. That is, work performance ratings are used by an organization for decision making (e.g., merit increases, promotions) in an imperfect form and, therefore, the effect sizes should not be corrected for unreliability. Consequently, the prescriptions provided by the present research for implementing performance ratings reflect the fact that ratings are used in imperfect form by organizations. On the other hand, the present research is also concerned with the development of performance rating theory, and from that perspective, the effect sizes should be corrected for unreliability. Such corrections provide the basis for estimating the theoretical influence of the study characteristics on the effect sizes. However, as will be discussed below, since the reliabilities of the effect sizes and study characteristics were quite high, the impact of the corrections would be minimal.

#### Analytic Approach

Hunter et al. (1982) emphasized correlation coefficients in their approach to meta-analysis whereas the present review employed ICCs as effect size estimates. However, ICCs can be treated as having a sampling distribution approximately the same as the Pearson product moment correlation (Kavanagh et al., 1971). Thus, the Hunter et al. sampling error formulas were adopted for use with ICCs.

As suggested by Hunter et al. (1982), the variance attributable to the artifact of sampling error was estimated with weighted means and variances. The

weighted mean ICC was computed for each of the dependent variables:

$$\overline{ICC} = \frac{\sum (N_1 \times ICC_1)}{\sum N_1}, \quad (1)$$

where the subscripting refers to each study's intraclass correlation and the number of observations used to calculate that correlation. The number of observations for convergent validity was the number of ratees; for method bias, the number of ratees times the number of methods; and for discriminant validity, the number of ratees times the number of traits. As shown in equation (2), the observed variance of the ICCs was computed as the weighted average squared errors:

$$s_{ICC}^2 = \frac{\sum (N_1 \times (ICC_1 - \overline{ICC})^2)}{\sum N_1}. \quad (2)$$

Next, the variance due to sampling error was computed:

$$s_e^2 = \frac{\sum \left[ \frac{N_1 \times (1 - ICC_1^2)}{N_1 - 1} \right]}{\sum N_1}. \quad (3)$$

The difference between the variance observed for the ICCs and the variance due to sampling error was used as an estimate of the true or population variance for ICCs:

$$s_T^2 = s_{ICC}^2 - s_e^2 . \quad (4)$$

The ratio of true variance to observed variance multiplied by 100% is the percent of observed variance that remains after correcting for sampling error. Naturally, if all the observed variance of the ICCs is attributable to sampling error, there can be no moderator variables. That is, the ICCs from the studies can be assumed to have the same expected value in the population.

Pearlman, Schmidt, and Hunter (1980) suggested that 25% be used as a rule-of-thumb regarding moderator variables. That is, if 25% or less of the observed variance remains after correcting for study artifacts, then it is inappropriate to search for moderator variables.

Table 7 presents the average ICC, observed variance, sampling error variance, true variance, and the percent of observed variance unexplained by sampling error for each of the three dependent variables. On the average, ratings exhibited fairly high convergent validity (.346), moderate method bias (.223), and relatively low discriminant validity (.128). As none of the dependent variable distributions violated the 25% rule, an examination for potential moderator variables was warranted. Subsequent analyses examined which variables influenced the degree of convergent validity, method bias, and discriminant validity exhibited in performance ratings.

#### Regression Analyses

A weighted least squares approach was used in the regression analyses to correct for sampling error variance (see Draper & Smith, 1981). Each ICC was assumed to be distributed with a different sampling error variance, as would be the case if moderators explained differences in the effect sizes. Each ICC

**Table 7.** Average ICC, Observed Variance, Sampling Error Variance, True Variance, and the Percent of Observed Variance Unexplained by Sampling Error for Each of the Dependent Variables

Dependent variable	$\overline{ICC}$	$s^2_{ICC}$	$s^2_e$	$s^2_T$	% unexplained
Convergent validity	.346	.01674	.00755	.00919	55
Method bias	.223	.01859	.00381	.01478	80
Discriminant validity	.128	.00753	.00117	.00636	84

and its corresponding study characteristics in an analysis were multiplied by the reciprocal of the square root of the appropriate sampling error variance, and the resulting values were subjected to ordinary regression (cf. Hedges, 1982).

Although corrections for unreliability are appropriate for theory development, they were not done in the present research. The rationale for this decision was based on a consideration of the underlying linear model for the corrections (Dickinson, 1985) and the magnitude of the reliabilities available for the corrections.

The study characteristics contained measurement error, because the values of the characteristics were determined by fallible coders. This measurement error due to coding may be estimated by two methods: interrater agreement (Jones et al., 1983; Shrout & Fleiss, 1979) or code-recode agreement with the same or different coders. Although these methods may tap different types of measurement errors, it is likely that both methods can be made to yield high reliabilities

through the care given to properly designing the coding procedures. In a regression analysis, adjustments can be made to correct for measurement errors due to coding. However, a recent empirical study (Dickinson, 1985) suggests that such corrections will have little effect if the predictors in the linear model have reasonably high reliabilities (e.g., .80 or higher).

The study effect sizes also contained measurement errors due to the manner in which the study was conducted. These errors limit the amount of variation in study effects that can be replicated (Mellenbergh, 1977). Thus, a method for estimating the reliability of the ICCs would be to redo the research studies and correlate the original and resulting ICCs.

Of course, replicability of the ICCs can also be understood by the test-retest reliability and intercorrelations of the measures that determine each study's ICCs. If the MTMM measures in a study have reasonable reliability, a study's ICCs will be quite reliable. Since ICCs are computed as linear combinations of the MTMM measures, a study's ICCs have reliabilities that are stepped up from the measures' average reliabilities with the Spearman-Brown formula. For example, assuming a conservative reliability of .60 for a rating scale (Schmidt & Hunter, 1977) and six scales in the MTMM correlation matrix, the reliability of the ICC for convergent validity would be .90. The reliabilities for method bias and discriminant validity would be greater.

In the present meta-analysis, the reliabilities of the study characteristics and the ICCs were sufficiently high to suggest that correcting for the influence of measurement error for theoretical purposes was not necessary.

Three stepwise regression analyses were conducted for each subset of study characteristics; one for each of the three dependent variables. Tables 8, 9, and 10 present the study characteristics that entered the regressions and their corresponding beta weights (i.e., standardized regression coefficients) for convergent validity, method bias, and discriminant validity,

**Table 8. Stepwise Regression by Subset for  
Convergent Validity**

Subsets of study characteristics	Beta weight
1. <u>Procedures to Develop Dimensions</u>	
C. Factor analysis	-.12
D. Expert prescription	.23
2. <u>Involvement in the Development of the Rating Scale</u>	
B. Expert involved	.14
D. Existing scale modified	-.21
3. <u>Content and Number of Dimensions</u>	
A. Strictly behavioral	.27
F. Number of ratings per dimension	-.16
4. <u>Rating Format</u>	
A. BARS/BES	.27
C. MSS/"other"	.15
5. <u>Rating Source</u>	
C. Peers	-.12
D. Self	-.14
E. Subordinates	-.14
F. Students	-.16
6. <u>Rater/Ratee Characteristics: Gender and Ratio</u>	
-None-	--
7. <u>Rating Context: Purpose Location, and Rater Training</u>	
D. Academia	-.20
8. <u>Study Design</u>	
A. Method: source vs. format	-.40

**Note.** Beta weights are reported only for variables that entered into the equations ( $p < .05$ ).

Table 9. Stepwise Regression by Subset for  
Method Bias

Subsets of study characteristics	Beta weight
<u>1. Procedures to Develop Dimensions</u>	
A. Systematic development	-.44
B. Derivation from existing scale	-.32
D. Expert prescription	-.23
<u>2. Involvement in the Development of the Rating Scale</u>	
A. Raters/ratees involved	-.22
<u>3. Content and Number of Dimensions</u>	
A. Dimensions strictly behavioral	-.41
F. Number of ratings per dimension	-.23
<u>4. Rating Format</u>	
A. BARS/BES	-.56
B. Graphic	-.45
C. MSS/other	-.31
<u>5. Rating Source</u>	
B. 2nd level supervisor	-.26
D. Self	.29
F. Students	-.32
<u>6. Rater/Ratee Characteristics: Gender and Ratio</u>	
C. All male ratees	.31
<u>7. Rating Context: Purpose Location, and Rater Training</u>	
A. Rating for non-administrative purposes	.25
F. Rater training	-.20

Table 9. (concluded)

Subsets of study characteristics	Beta weight
<u>8. Study Design</u>	
A. Method: source vs. format	.41
B. Mixed vs. nested/crossed	-.34

Note. Beta weights are reported only for variables that entered the equations ( $p < .05$ ).

respectively. For convergent validity, 14 study characteristics entered significantly ( $p < .05$ ); for method bias, 17; and for discriminant validity, 7. Thus, a total of 38 characteristics entered the regression equations out of 111 possible entries.

The beta weights are presented to indicate the direction of the relationships. The weighted correlations between the study characteristics and the dependent variables are not reported, because they provide misleading information about the magnitude and direction of the relationships. Unweighted correlations are reported in Table 6. The nature of the relationships are best described by examining the weighted means presented in the subgroup analyses.

#### Subgroup Analyses

For each study characteristic that was identified by the regression analyses, a subgroup analysis was conducted. These analyses served two purposes: (a) They provided an additional rigorous test of the study characteristics to minimize findings that capitalized on chance, and (b) they described the magnitude of the relationship between a study characteristic and rating quality by examining the subgroup weighted means.

For each study characteristic, with the exception of the one significant continuous variable, the studies were partitioned into two groups: those studies that exhibited the study characteristic (the "yes" group) and those studies that did not (the "no" group). For the

Table 10. Stepwise Regression by Subset for  
Discriminant Validity

Subsets of study characteristics	Beta weight
1. <u>Procedures to Develop Dimensions</u>	
-None-	--
2. <u>Involvement in the Development of the Rating Scale</u>	
-None-	--
3. <u>Content and Number of Dimensions</u>	
F. Number of ratings per dimension	.52
4. <u>Rating Format</u>	
-None-	--
5. <u>Rating Source</u>	
E. Subordinates	-.16
F. Students	.45
6. <u>Rater/Ratee Characteristics: Gender and Ratio</u>	
D. All female ratees	-.22
7. <u>Rating Context: Purpose, Location, and Rater Training</u>	
D. Academia	.50
F. Rater training	.23
8. <u>Study Design</u>	
B. Mixed vs. nested/crossed	.42

Note. Beta weights are reported only for variables  
that entered the equations ( $p < .05$ ).

yes and no groups, their weighted mean, observed variance, variance due to sampling error, true variance, and the percent of observed variance unexplained by sampling error were computed.

These analyses are similar to those reported by Hunter et al. (1982), with one important distinction. In Hunter et al., the study characteristics were uncorrelated. In the current research, the study characteristics were intercorrelated response categories. For example, in some studies, first-level supervisors, peers, and subordinates may all have provided ratings. Therefore, it was not possible to categorize studies into those that used supervisor ratings versus those that used peer ratings. For this reason, each study characteristic was considered separately.

It is important to note that separating the studies on the basis of a single study characteristic creates one subgroup of studies that share the characteristic in common and another subgroup that includes studies that exhibit all remaining characteristics (e.g., the set of studies that did not include peer ratings would have had the remaining sources of ratings collapsed together). Consequently, if the study characteristic is a moderator variable, the yes group should exhibit reduced variance but the no group, which could be highly heterogeneous, need not show reduced variability.

According to Hunter et al. (1982), if a study characteristic is a moderator variable, the subgroups partitioned on that characteristic should be able to pass two tests: (a) there should be mean differences between the subgroups, and (b) the subgroups should demonstrate true variances that are smaller than the true variance of the total group. The second test was modified for the current research because of the potential heterogeneity of studies in the no group.

In sum, three conditions needed to be met for a study characteristic to be considered a moderator variable. It must: (a) enter significantly in the regression analyses, (b) demonstrate mean differences between the yes and no subgroups, and (c) demonstrate a reduction in the true variance of the yes group compared to the true variance of the total group of studies.

Those study characteristics that entered into the regressions but did not demonstrate mean differences were acting as suppressor variables. Because suppressor variable effects in psychological research are rarely replicable, those results are not discussed further.

Tables 11, 12, and 13 show the subgroup analyses. For comparison purposes, the tables also include the relevant statistics for the total group of studies. Eight study characteristics failed to meet the conditions established for a subgroup analysis. These characteristics are noted in the tables. The final number of significant pairs of study characteristics and dependent variables was 30.

#### IV. DISCUSSION

A variety of study characteristics were moderator variables. These are discussed by subsets of characteristics. On the basis of the results, several prescriptive recommendations are offered to improve MTMM properties. These are described by property (i.e., convergent validity, method bias, and discriminant validity).

The quantitative review technique identified gaps in the literature and deficiencies in the reporting of methodology and results. These omissions, along with the results of the meta-analysis, pointed to several specific research questions. These questions are posed to provide direction for future R&D efforts.

##### Developmental Procedures

Several procedures used to develop performance dimensions influenced the amount of convergent validity and method bias. The use of expert prescriptions as a component of the developmental process produced greater convergent validity and less method bias. However, the use of expert prescriptions was usually accompanied by rater/ratee involvement in the developmental procedure ( $r = .45$ ). This involvement is probably a desirable strategy. When expert prescriptions are used, particularly if the experts are from outside the organization, raters and/or ratees should be involved to ground the dimensions in organizational reality.

Table 11. Subgroup Analyses for Convergent Validity

Moderators (by subset)	Number of matrices	$\overline{ICC}$	$S^2_{ICC}$	$S^2_e$	$S^2_T$	% unex- plained
<u>Procedures to Develop Dimensions</u>						
Factor analysis?						
Yes	7	.243	.01070	.00594	.00476	44
No	24	.391	.01267	.00828	.00439	35
Expert prescriptions?						
Yes	5	.480	.00428	.00353	.00075	18
No	26	.289	.01118	.00926	.00192	17
<u>Involvement in the Development of the Rating Scale</u>						
Experts involved?						
Yes	6	.441	.00909	.00467	.00442	49
No	25	.305	.01441	.00879	.00562	39
Existing scale modified?						
Yes	7	.218	.00609	.00663	-	0
No	24	.398	.01174	.00793	.00381	32
<u>Content and Number of Dimensions</u>						
Behavioral?						
Yes	19	.406	.01416	.00794	.00622	44
No	12	.269	.00952	.00707	.00245	26
<u>Rating Format</u>						
BARS/BES?						
Yes	8	.461	.00337	.00492	-	0
No	23	.288	.01343	.00889	.00454	34
MSS/Other? <sup>a</sup>						
Yes	8	.399	.01856	.00729	.01127	61
No	23	.328	.01483	.00764	.00719	48

Table 11. (concluded)

Moderators (by subset)	Number of matrices	$\overline{ICC}$	$s^2_{ICC}$	$s^2_e$	$s^2_T$	% unex- plained
<u>Rating Source</u>						
Peers?						
Yes	10	.294	.00751	.01146	-	0
No	21	.361	.01082	.00641	.01201	65
Self?						
Yes	10	.219	.00751	.01009	-	0
No	21	.394	.01206	.00660	.00546	45
Subordinates?						
Yes	4	.288	.00134	.00699	-	0
No	27	.356	.01874	.00765	.01109	59
Students?						
Yes	4	.150	.00206	.00827	-	0
No	27	.377	.01196	.00744	.00452	38
<u>Rating Context</u>						
Academic?						
Yes	6	.175	.00550	.01030	-	0
No	25	.379	.01220	.00703	.00517	42
<u>Study Design</u>						
Method in MTMM?						
Source	28	.289	.01082	.00729	-	0
Format	3	.498	.00089	.00187	-	0
Total Group	31	.346	.01674	.00755	.00919	55

<sup>a</sup>Variable eliminated because it failed to demonstrate a reduction in the true variance of the yes group compared to that of the total group.

Table 12. Subgroup Analyses for Method Bias

Moderators (by subset)	Number of matrices	$\overline{ICC}$	$S^2_{ICC}$	$S^2_e$	$S^2_T$	% unex- plained
<u>Procedures to Develop Dimensions</u>						
Systematic?						
Yes	14	.142	.00595	.00393	.00202	34
No	17	.294	.01886	.00370	.01516	80
Derivation from existing scale?						
Yes	11	.231	.01289	.00377	.00912	71
No	20	.219	.02188	.00382	.01806	83
Expert prescription?						
Yes	5	.091	.00297	.00241	.00056	19
No	26	.274	.01522	.00434	.01088	71
<u>Involvement in Development</u>						
Raters/ratees?						
Yes	6	.149	.00592	.00279	.00312	53
No	25	.252	.02052	.00420	.01632	80
<u>Content and Number of Dimensions</u>						
Behavioral?						
Yes	19	.153	.00681	.00471	.00210	31
No	12	.302	.01993	.00280	.01713	86
<u>Rating Format</u>						
BARS/BES?						
Yes	8	.116	.00065	.00357	-	0
No	23	.269	.01927	.00390	.01537	80
Graphic? <sup>a, b</sup>						
Yes	15	.236	.01804	.00383	.01421	79
No	16	.211	.01880	.00378	.01502	80
MSS/Other?						
Yes	8	.157	.01017	.00412	.00605	59
No	23	.246	.01941	.00370	.01571	81

Table 12. (continued)

Moderators (by subset)	Number of matrices	$\overline{ICC}$	$S^2_{ICC}$	$S^2_e$	$S^2_T$	% unex- plained
<u>Rating Source</u>						
2nd level?						
Yes	12	.147	.00297	.00451	-	0
No	19	.264	.02213	.00343	.01870	85
Self?						
Yes	10	.284	.02052	.00425	.01627	79
No	21	.200	.01588	.00363	.01225	77
Students?						
Yes	3	.149	.00839	.00436	.00403	48
No	27	.234	.02022	.00373	.01649	82
<u>Rater/Ratee Characteristics</u>						
All male ratees? <sup>b</sup>						
Yes	3	.306	.02662	.00437	.02225	84
No	27	.201	.01343	.00387	.00956	71
<u>Rating Context</u>						
Non-administrative? <sup>b</sup>						
Yes	6	.313	.03490	.00360	.03130	90
No	25	.203	.01268	.00385	.00883	70
Rater training?						
Yes	3	.090	.00930	.00453	.00477	51
No	28	.236	.01761	.00373	.01388	79
<u>Study Design</u>						
Method in MTMM?						
Source	28	.266	.01627	.00447	.01180	73
Format	3	.088	.00165	.00170	-	0

Table 12. (concluded)

Moderators (by subset)	Number of matrices	$\overline{ICC}$	$S^2_{ICC}$	$S^2_e$	$S^2_T$	% unex- plained
Fully mixed?						
Yes	17	.160	.01010	.00349	.00661	65
No	14	.326	.01541	.00432	.01109	72
Total Group	31	.223	.01859	.00381	.01478	80

<sup>a</sup>Variable eliminated because it failed to demonstrate mean differences between the yes and no subgroups.

<sup>b</sup>Variable eliminated because it failed to demonstrate a reduction in the true variance of the yes group compared to that of the total group.

The use of performance dimensions that were developed by factor analysis was negatively related to convergent validity. Factor analysis yields empirically distinct performance dimensions such that ratees should be ranked differently on each of the dimensions. Consequently, when ratings are averaged across dimensions to determine convergent validity, overall ratee differences are minimized. This finding does not suggest that the use of factor analysis should be avoided in developing performance dimensions. However, for situations where convergent validity is important (e.g., an average rating for administrative decisions), the use of factor analysis to specify dimensions may be problematic.

The use of systematic procedures to identify dimensions (e.g., use of job descriptions, discussions with subject-matter experts, surveys, retranslation) is negatively related to method bias. Not surprisingly, the use of systematic procedures is often accompanied by the use of behavioral dimensions ( $r = .72$ ), example-anchored scales ( $r = .65$ ), and dimensions that contain specific content ( $r = .60$ ). Perhaps the use of a systematic procedure by itself reduces method bias,

Table 13. Subgroup Analyses for Discriminant Validity

Moderators (by subset)	Number of matrices	$\overline{ICC}$	$S^2_{ICC}$	$S^2_e$	$S^2_T$	% unex- plained
<u>Rating Source</u>						
Subordinates?						
Yes	4	.058	.00058	.00076	-	0
No	27	.146	.00773	.00127	.00646	84
Students?						
Yes	4	.295	.00069	.00097	-	0
No	27	.101	.00333	.00120	.00213	64
<u>Rater/Ratee Characteristics</u>						
All female ratees? <sup>a</sup>						
Yes	6	.069	.00784	.00153	.00631	80
No	24	.142	.00785	.00109	.00676	86
<u>Rating Context</u>						
Academic setting?						
Yes	6	.280	.00341	.00141	.00200	59
No	25	.101	.00336	.00113	.00223	66
Rater training?						
Yes	3	.208	.00484	.00124	.00360	74
No	28	.121	.00712	.00116	.00596	84
<u>Study Design</u>						
Fully mixed? <sup>a</sup>						
Yes	17	.153	.00714	.00094	.00620	87
No	14	.075	.00418	.00166	.00252	60
Total Group	31	.128	.00753	.00117	.00636	84

<sup>a</sup>Variable eliminated because it failed to demonstrate a reduction in the true variance of the yes group compared to that of the total group.

regardless of the end product (i.e., specific, behavioral dimensions rated with example-anchored scales). Unfortunately, it is not possible to distinguish the contribution to reduced method bias of the procedure from its product.

#### Involvement in Development

Although some developmental procedures require the involvement of certain individuals (e.g., retranslation requires rater involvement), the participation of others is optional. Using retranslation as an illustration, raters must be involved, but ratees and experts may help raters with the development of the rating system. Therefore, involvement is considered separately from developmental procedures.

Involvement by raters and/or ratees is associated with reduced method bias, whereas involvement by experts is related to higher convergent validity. Furthermore, the use of modifications of existing scales is negatively related to convergent validity. Since existing scales are usually modified via factor analysis ( $r = .82$ ), raters or ratees are not involved in development. The data suggest that this is not a recommended strategy. Raters, ratees, and experts should all be involved to enhance the quality of ratings.

#### Content and Number of Dimensions

The use of behavioral dimensions was associated with higher convergent validity and lower method bias. No significant results were found for studies that used both behavioral and trait dimensions or those that used trait-oriented dimensions. The behavioral scales used in the MTMM studies were more likely to employ a BARS/BES or MSS/other format ( $r = .47$  and  $.32$ , respectively) than a graphic format ( $r = -.42$ ). The developmental procedure associated with BARS/BES ( $r = .65$ ) and MSS scales ( $r = .50$ ) tends to be more thorough than those used for graphic scales ( $r = -.75$ ). Therefore, the possibility remains that trait-oriented scales, when they are job-related and carefully developed, may yield high-quality ratings. However, as they are usually developed, behavioral scales demonstrate higher-quality ratings than trait-oriented scales.

More specific dimension content was associated with lower method bias. However, specific content did not enter the regression for method bias, perhaps because of its correlation with the behavioral dimensions predictor ( $r = .53$ ).

The greater the number of ratings per dimension, the lower the method bias and the greater the discriminant validity. This occurred despite the fact that BARS/BES scales tend to have fewer ratings per dimension than graphic scales. An outcome associated with greater ratings per dimension was lower convergent validity. Apparently, the additional ratings per dimension helped raters focus on ratee differences and increased their ability to discriminate among ratees.

Unfortunately, no conclusions can be drawn about specific types of dimension content. Questions remain regarding the quality of ratings based on an interpersonal skill dimension compared to a technical skill dimension, as well as who can best provide these ratings.

#### Rating Format

Example-anchored formats demonstrated high convergent validity and low method bias, and MSS/other formats exhibited low method bias. As stated earlier, it is difficult to separate a developmental procedure from its products. Only three studies used format as the method in the MTMM matrix. Additional research needs to examine the effects of different developmental procedures independent of rating format.

In a review of BARS/BES scales, Kingstrom and Bass (1981) concluded that there was little difference between the behaviorally anchored and other formats. However, they used a narrative review technique. On the basis of this meta-analysis, it is clear that BARS/BES and MSS formats yielded higher-quality ratings (i.e., greater convergent validity and/or lower method bias) than did the graphic format. It is unclear if the quality of ratings obtained with the graphic format could be improved by the use of a systematic developmental procedure which involves raters, ratees, and experts and which focuses on specific dimensions.

### Rating Source

Rating sources exhibited mixed findings. Peer, self-, subordinate, and student ratings were all associated with lower convergent validity; self-ratings were related to greater method bias; and subordinate ratings had low discriminant validity. On the other hand, student ratings demonstrated low method bias and high discriminant validity, and second-level supervisory ratings exhibited low method bias.

Taken as a whole, the findings indicate that different rating sources have different perspectives of ratee performance. This viewpoint is also supported by the finding that studies using source of ratings as the method in the MTMM matrix demonstrated much lower convergent validity and greater method bias than did those studies that focused on format. This viewpoint has important implications for the design of a performance rating system. If peers, supervisors, and subordinates observe work performance under different circumstances or even perceive the same performance differently, their separate perspectives of the ratees' performance provide unique information. Multiple rating sources may be necessary to be able to measure all aspects of work performance. An important step toward improving rating quality is to begin to identify which rating sources provide high-quality ratings on which dimensions. For example, supervisors may provide better ratings on technical dimensions, and peers may provide useful additional information on interpersonal dimensions.

The high method bias associated with self-ratings suggests that self-ratings either measure a different aspect of work performance than do other sources or they are inaccurate, perhaps due to inflation.

Student ratings, although low in convergent validity, demonstrated positive findings for method bias and discriminant validity. This was true even though student ratings often used existing scales ( $r = .43$ ) and the rating systems were less likely to have used a systematic developmental procedure ( $r = -.35$ ).

Student rating quality may be due to the purpose of these ratings and the typical student/instructor relationship. Student ratings are usually for feedback purposes; the ratings inform instructors how they have

performed on a variety of dimensions. In addition, student ratings are usually anonymous, and the student does not have to feed back the ratings directly to the instructor. Students may perceive honest ratings as being useful to the instructor, and they know that they can avert the discomfort of face-to-face interaction for feedback. This is a different situation than most other rating contexts. Ratings may be anonymous (e.g., for research purposes), and they may be used for feedback (e.g., for growth and development), but to the authors' knowledge, ratings are seldom both anonymous and used for feedback. These results for student ratings suggest that future research should examine the joint effects of purpose of rating and anonymity.

#### Rater/Ratee Characteristics

None of the rater/ratee characteristics examined in the meta-analysis were significantly related to rating quality. This may be attributable to the paucity of rater/ratee information reported in the research studies. Of the many variables of interest, only rater/ratee sex information could be analyzed and, even then, only a limited number of studies contained the necessary information.

#### Rating Context: Purpose, Location, and Training

Some research suggests that rating quality may be affected by rating purpose (e.g., McIntyre, Smith, & Hassett, 1984). The present results indicated that MTMM properties were not related to rating purpose. However, only seven MTMM studies reported the purpose for rating.

Ratings made in an academic location demonstrated high discriminant validity but low convergent validity. This is the only significant finding related to rating location and may be attributable to the special circumstances mentioned earlier (i.e., rating for feedback purposes and rater anonymity). The failure to find other effects suggests that rating quality is not moderated by location.

Rater training was associated with low method bias and greater discriminant validity. These findings are extremely encouraging, particularly since discriminant validity was typically low and not related to many variables that have potential for manipulation. Unfortunately, only three studies provided information

on rater training, and it was not possible to identify which type of training was most effective.

### Prescriptive Recommendations

On the basis of the results of the present review, several prescriptive recommendations are offered to help improve each property. Naturally, rating quality is not the only important feature of a rating system. Other considerations must include the cost, legality, and acceptability of the system.

To improve convergent validity, which is particularly important for those rating systems that are used only for administrative decisions:

1. Use behavioral dimensions. Trait dimensions may be useful if they are properly developed and if they are clearly job-relevant. However, systems that use trait dimensions, as they are typically developed, demonstrate lower convergent validity than do those that use behavioral dimensions.

2. Use example-anchored scales. BARS and BES formats demonstrate greater convergent validity. This may be due to the developmental procedure that is followed and not simply the rating format.

3. Avoid using factor analysis as the primary procedure for defining performance dimensions.

4. Involve experts in the development of the rating system, particularly in the identification of dimensions. In some studies, experts performed a qualitative clustering on the information gathered from employees. Experts may be organizational members or external consultants; however, experts should not be the only participants in development. Those who will be using the system should be involved, to ground it in organizational reality.

5. Do not modify existing rating scales. When existing scales are modified and adapted, convergent validity is low. If possible, rating scales which incorporate the above recommendations should be developed locally.

To reduce method bias:

1. Use systematic developmental procedures. This includes job descriptions, surveys, and retranslation. The optimal procedure cannot be recommended on the basis of available research.

2. Involve experts in the development of dimensions. (See recommendation four under convergent validity.)

3. Involve raters/ratees in the development of dimensions. This may also improve the acceptability of the rating system.

4. Use behavioral dimensions. (See recommendation one under convergent validity.)

5. Use example-anchored scales. (See recommendation two under convergent validity.) MSS scales also demonstrate lower method bias, but the effect is somewhat smaller.

6. Provide rater training. Studies that provided rater training demonstrated considerably less method bias. Specific training content and methods cannot be recommended on the basis of the available MTMM studies.

To improve discriminant validity, which is particularly important for those systems that feed back ratings for employee growth and development:

1. Provide rater training. (See recommendation six under method bias.) It is speculated that if the purpose of rating is direct feedback, then rater training should incorporate a component on how to give feedback. Raters may be reluctant to provide low ratings on some dimensions, because they are unable to give negative feedback. This would reduce discriminability across dimensions. As students do not have to provide direct feedback, their ratings demonstrate greater discriminant validity.

2. Use scales with several ratings per dimension.

3. Collect ratings from multiple sources. Different sources have different perspectives of ratee work performance and can provide useful information for feedback purposes.

## Research Questions

The quantitative review technique provided an identification of gaps in the literature, as well as deficiencies in the reporting of methodology and results. These gaps and deficiencies, together with the findings from the meta-analysis, point to specific research questions:

### 1. What is the relative influence of the developmental procedure versus its product?

Several reviewers (Landy & Farr, 1980; Schwab, Heneman, & DeCotiis, 1975) have suggested that format research is passé, but this meta-analysis provided evidence that BARS/BES demonstrate low method bias and high convergent validity. These findings may be because BARS/BES formats are example-anchored scales and have specific content, or because the formats are developed through retranslation which incorporates the participation of raters, ratees, and even experts. If the developmental procedure is the key, then graphic scales may also be a useful format.

### 2. Which rater sources are in the best position to evaluate which performance dimensions?

Do supervisors provide better-quality ratings for technical performance dimensions than do peers? Do peers provide useful ratings for interpersonal performance dimensions? What is the role of "opportunity to observe"? To address these issues, the MTMM design must be extended to include multiple formats, multiple sources, and different types of dimensions (e.g., technical, interpersonal). Analyses must assess their interactions.

### 3. Does rating purpose affect rating quality?

Do raters provide higher-quality ratings for research purposes than for administrative purposes? Does a rater whose ratings affect someone's future (i.e., pay, promotion) provide ratings with reduced discriminant validity? When ratings are for feedback purposes, does the need to provide the ratee face-to-face feedback affect rating quality? What is the influence of rater anonymity on rating quality? Future research should be designed that varies purposes and anonymity and assesses rating quality.

4. Which types of rater training yield higher-quality ratings?

Rater training is linked to greater discriminant validity and lower method bias. What should be the content of the training? Should training focus on the rating scales per se, performance standards, observation skills, or some combination of these? What training methods work best (e.g., group discussion, videotapes, lecture)? What is the effect of training content on convergent validity, method bias, and discriminant validity? Data should be collected before and after training to evaluate changes in MTMM properties.

5. Are rater/ratee characteristics related to rating quality?

In psychology, almost every area of research goes through a "trait stage." The authors are not optimistic about the value of research in this area. However, reporting deficiencies in the previous research studies made it impossible to assess the effects of rater/ratee characteristics on MTMM properties.

## V. CONCLUSIONS

A good integration of research studies shows how much is known in an area, but it also shows how little is known. "In this sense, it is only the beginning" (Feldman, 1971, p. 100). The prescriptive recommendations that were derived from the analysis of MTMM properties indicate that much is known. The research questions that were derived from this analysis and the gaps in the literature indicate how little is known and point to research needs.

Multitrait-multimethod research provides valuable information about rating quality that cannot be collected through other approaches. This information is essential to improving performance rating systems. Previous reviews of performance ratings have focused on the psychometric properties of ratings. This review identified variables that influence convergent validity, method bias, and discriminant validity. What remains to be done is to apply this information and to continue to conduct MTMM research so as to better understand how ratings are made and how they can be improved.

## REFERENCES

- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 19, 3-11.
- Borich, G. D., Malitz, D., & Kugle, C. L. (1978). Convergent and discriminant validity of five classroom observation systems: Testing a model. Journal of Educational Psychology, 70, 119-128.
- Borman, W. C., & Rosse, R. L. (1978). Format and training effects on rating accuracy and rater errors. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Bullock, R. J., & Svyantek, D. J. (1985). Analyzing meta-analysis: Potential problems, an unsuccessful replication, and evaluation criteria. Journal of Applied Psychology, 70, 108-115.
- Burton, N. W. (1981). Estimating scorer agreement for nominal categorization systems. Educational and Psychological Measurement, 41, 953-962.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provisions for scaled disagreement or partial credit. Psychological Bulletin, 70, 213-220.
- DeCotiis, T. A., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. Academy of Management Review, 3, 635-646.
- Dickinson, T. L. (1985, August). The general linear model and errors of measurement. Paper presented at the meeting of the American Psychological Association, Los Angeles, CA.

- Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. Journal of Applied Psychology, 65, 147-154.
- Draper, N. R., & Smith H. (1981). Applied regression analysis (2nd ed.). New York: Wiley & Sons.
- Feldman, K. A. (1971). Using the work of others: Some observations on reviewing and integrating. Sociology of Education, 44, 86-102.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage Publications.
- Hedges, L. V. (1982). Fitting continuous models to effect size data. Journal of Educational Statistics, 7, 245-270.
- Hull, C. H., & Nie, N. H. (1981). SPSS update 7-9. New procedures and facilities for release 7-9. New York: McGraw-Hill.
- Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage Publications.
- Jackson, G. B. (1980). Methods for integrative reviews. Review of Educational Research, 50, 438-460.
- Jacobs, R., Kafry, D., & Zedeck, S. (1979). Consistency in multidimensional performance evaluations: An analysis of raters and dimensions. Catalog of Selected Documents in Psychology, 9, 25. MS 1834.
- Jenkins, G. D., Nadler, D. A., Lawler, E. E., III, & Cammann, C. (1975). Standardized observations: An approach to measuring the nature of jobs. Journal of Applied Psychology, 60, 171-181.
- Jones, A. P., Johnson, L. A., Butler, M. C., & Main D. S. (1983). Apples and Oranges: An empirical comparison of commonly used indices of interrater agreement. Academy of Management Journal, 26, 507-519.

- Kane, J. S., & Lawler, E. E., III. (1978). Methods of peer assessment. Psychological Bulletin, 85, 555-586.
- Kavanagh, M. J., Borman, W. C., Hedge, J. W., & Gould, R. B. (1986). Job performance measurement classification scheme for validation research in the military (AFHRL-TP-85-51, AD-A164837). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 75, 34-49.
- Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. Personnel Psychology, 34, 263-289.
- Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. Journal of Applied Psychology, 70, 56-65.
- Landy, F. J., & Farr, J. L. (1980). Performance rating. Psychological Bulletin, 87, 72-107.
- Lawler, E. E., III. (1967). The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 51, 369-381.
- Mathieu, J., & Tannenbaum, S. I. (1983, March). Analyzing the analyses today. In S. I. Tannenbaum & E. Salas (Co-chairs), Meta-analysis in I/O and O.B.: Probes of the technique, current practices, and the future. Symposium presented at the I/O and O. B. Graduate Student Convention, Chicago, IL.
- McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.
- Mellenbergh, G. J. (1977). The replicability of measures. Psychological Bulletin, 84, 378-384.

- Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 65, 373-406.
- Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88, 413-428.
- Schmidt, F. L., & Hunter J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.
- Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.
- Schwab, D. P., Heneman, H. G., III, & DeCotiis, T. (1975). Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 28, 549-562.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.
- Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.
- Wheeler, A. E., & Knoop, H. R. (1982). Self, teacher and faculty assessments of student tracking performance. Journal of Educational Research, 75, 178-181.

## APPENDIX A: ANNOTATED BIBLIOGRAPHY

This is an annotated bibliography of all studies included in the meta-analysis. All studies reported either a multitrait-multimethod correlation matrix or an ANOVA summary table. For studies that reported the correlation matrix, the intraclass correlations (ICCs) were recomputed according to Bartko's (1966) definition (i.e., the ratio of a source's variance component to the sum of all relevant variance components). For studies that did not report the correlation matrix, the ICCs were recomputed using the mean squares reported in the summary table. The ICCs were described verbally as follows: high, good (above .30), medium, moderate (.20 to .29), and low, poor (less than .20). The variance component (VC) for error was described as: high (above .30), moderate (.20 to .29), and low (less than .20).

A number of studies included "overall performance" as one of the rating dimensions. In these instances, the overall performance dimension was eliminated from the calculations. This was done because overall performance ratings would be highly correlated with ratings on the other dimensions and would spuriously affect findings of convergent validity and discriminant validity.

In some studies, dimensions other than overall performance were eliminated. These dimensions were eliminated because they measured job attitudes and not work performance. Examples of eliminated dimensions are "satisfaction with supervisors" and "job satisfaction."

Due to incomplete or ambiguous reporting in some studies, inferences may have been made by the present authors (e.g., sample sizes). Interested persons may contact the authors of the present study for copies of the data used in the meta-analysis.

### Reference

- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 19, 3-11.

### Annotated Bibliography

- Arvey, R. D., & Hoyle, J. C. (1974). A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. Journal of Applied Psychology, 59, 61-68.

One hundred three supervisors rated 200 systems analysts employed by a large midwestern computer manufacturing company. Two rating scales were developed in a series of workshops with appropriate personnel. Workshop participants were asked to consider the major performance dimensions of the jobs in question. Eleven dimensions were identified. Participants then generated specific behavioral incidents to serve as anchor points for each dimension. The first rating scale was a continuum in which behavioral incidents were assigned a value from 7 (highly effective) to 1 (highly ineffective). The second scale consisted of the behavioral incidents with no values assigned. Raters indicated whether the ratee was better than, the same as, or not as good as the individual described in the incidents. Analysis of the multitrait-multimethod correlation matrix indicated good convergent validity ( $ICC = .498$ ,  $p < .001$ ), low discriminant validity ( $ICC = .114$ ,  $p < .001$ ), low method bias ( $ICC = .085$ ,  $p < .001$ ), and high error variance ( $VC = .32$ ). (19 references)

- Baird, L. S. (1977). Self and superior ratings of performance: As related to self-esteem and satisfaction with supervision. Academy of Management Journal, 20, 291-300.

Employees in a large state agency participated in this study. One hundred sixty-five participants rated themselves. These ratees were also rated by their superiors, using the same rating scale. All participants were told that the ratings would be used for research purposes only and would remain confidential. The a priori rating scale developed for this study consisted of a comparative rating that evaluated the ratee on four dimensions of job performance, relative to all other employees who reported to the same supervisor. Results indicated good convergent validity ( $ICC = .352$ ,  $p < .001$ ). There was also low discriminant validity ( $ICC = .026$ ,  $p < .001$ ), a high degree of method bias ( $ICC = .515$ ,

$p < .001$ ), and low error variance ( $VC = .15$ ) in the ratings. (22 references)

Blackburn, R. T., & Clark, M. J. (1975). An assessment of faculty performance: Some correlates between administrator, colleague, student, and self-ratings. Sociology of Education, 48, 242-256.

Forty-five full-time faculty members at a midwestern college rated themselves on two dimensions of faculty performance. Ratees were also rated by every other teacher in his/her curricular division and by the college's administrators. The ratees understood that this study would be used for employee growth and development. Ratings from the three sources were made on two 5-point scales relating to the two dimensions. Results indicated high convergent validity ( $ICC = .335$ ,  $p < .001$ ) and low discriminant validity ( $ICC = .123$ ,  $p < .001$ ). There was also a moderate degree of method bias ( $ICC = .282$ ,  $p < .001$ ) and high error variance ( $VC = .32$ ). (42 references)

Borman, W. C. (1974). The rating of individuals in organizations: An alternative approach. Organizational Behavior and Human Performance, 12, 105-124.

The ratees in this study were 27 secretaries from five academic departments at a large midwestern university. They were rated by one or more instructors to whom they were assigned. Each secretary's work was also rated by those peers with whom they came into day-to-day contact. Although supervisors and peers developed their own group's behavioral expectation scale (BES), all raters used both scales. Each secretary was rated on seven performance dimensions. There was evidence of high convergent validity ( $ICC = .312$ ,  $p < .001$ ). Support for discriminant validity was low ( $ICC = .077$ ,  $p < .031$ ), and there was little evidence of method bias ( $ICC = .171$ ,  $p < .001$ ). There was also high error variance ( $VC = .51$ ). (16 references)

Borman, W. C., Hough, L. M., & Dunnette, M. D. (1976). Development of behaviorally based rating scales for evaluating the performance of U.S. Navy recruiters (N-TR-76-31). San Diego, CA: Navy Personnel Research and Development Center.

Twenty-four Navy recruiters and three supervisors participated in this study. The recruiters were from eight recruiting stations in the Minneapolis/St. Paul area. Each recruiter rated himself and the one to three other recruiters serving the same station. The chief recruiter and a zone supervisor rated 14 of the recruiters, and the enlisted programs officer (EPO) rated the 10 remaining recruiters. Thus, there were self-, peer, and superior ratings. The Behavior Summary Scales were developed in a series of workshops involving recruiters, recruiting supervisors, and subject-matter experts. Examples of effective and ineffective performance were generated, and these examples were tentatively divided into performance dimensions. The dimensions were discussed and revised. At this stage, the examples were assigned to the dimensions and scaled. The examples retained were grouped together to provide a definition for each of the eight dimensions. The Behavior Summary Scales were then placed on a 10-point rating scale ranging from "ineffective performance" to "extremely effective performance." Analysis of the multitrait-multimethod correlation matrix indicated moderate convergent validity ( $ICC = .233$ ,  $p < .001$ ) and low discriminant validity ( $ICC = .054$ ,  $p < .008$ ). However, there was also moderate evidence of method bias ( $ICC = .283$ ,  $p < .001$ ) and high error variance ( $VC = .50$ ). (24 references)

Boruch, R.F., Larkin, J. D., Wolins, L., & MacKinney, A. C. (1970). Alternative methods of analysis: Multitrait-multimethod data. Educational and Psychological Measurement, 30, 833-853.

Study 1. The "most effective" and "least effective" subordinates of 111 supervisors of production and management operations in a large American corporation rated their supervisors. The subordinates were asked to describe their supervisors on four dimensions of job performance. The rating instrument was a continuum ranging from "1" (very nondescriptive) to "99" (very

descriptive). The designations of "most effective" and "least effective" were provided by the supervisors. Analysis of the multitrait-multimethod correlation matrix indicated a moderate degree of convergent validity ( $ICC = .246, p < .001$ ) and low discriminant validity ( $ICC = .110, p < .001$ ). However, there was also moderate evidence of method bias ( $ICC = .287, p < .001$ ) and high error variance ( $VC = .43$ ) in the ratings.

Study 2. One hundred twenty-four department heads from 24 industrial installations within an American corporation were rated by their superiors, their most effective subordinates, and their least effective subordinates. The department heads were rated on six dimensions of personal traits, using a rating instrument which was designed for a previous study. Results indicated that there was high convergent validity ( $ICC = .337, p < .001$ ) and low discriminant validity ( $ICC = .035, p < .001$ ). There was also high method bias ( $ICC = .397, p < .001$ ) and moderate error variance ( $VC = .27$ ). (23 references)

Braskamp, L. A., Caulley, D., & Costin, F. (1979). Student ratings and instructor self-ratings and their relationship to student achievement. American Educational Research Journal, 16, 295-306.

Study 1. Nineteen graduate teaching assistants in charge of sections of a one-semester course in psychology offered in a large midwestern university served as the ratees. Students in these courses rated their instructors on five dimensions of teacher performance. The five dimensions consisted of an average of 4.8 separate items. The rating forms were completed within the last two weeks of the semester. At the end of the semester prior to receiving the students' ratings, the instructors rated themselves using the same rating scale as the students. Analysis of the multitrait-multimethod correlation matrix revealed moderate convergent validity ( $ICC = .217, p < .001$ ). The analysis also indicated low discriminant validity ( $ICC = .146, p < .02$ ) and method bias ( $ICC = .176, p < .001$ ). There was high error variance ( $VC = .55$ ).

Study 2. This study was identical to the study described above, except that it was carried out in a different semester; and the number of ratees was 17.

The results of this study were, however, considerably different from those of the study described above. Analysis of the correlation matrix again indicated high convergent validity ( $ICC = .343$ ,  $p < .001$ ). The results also indicated moderate discriminant validity ( $ICC = .238$ ,  $p < .001$ ) and no method bias ( $ICC = .009$ ,  $p < .371$ ). There was high error variance ( $VC = .46$ ). (15 references)

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development of behaviorally based rating scales. Journal of Applied Psychology, 57, 15-22.

In this study, 527 department managers in retail stores throughout the United States were rated by their store managers. There were two rating methods used: behavioral expectation scales and summated graphic scales. These rating scales were developed in a series of workshops. In the first workshop, participants were asked to write at least five effective and five ineffective critical incidents of department manager performance. These were then subjected to a qualitative cluster analysis, and definitions of the categories were written. The dimension definitions were discussed by the participants and were adjusted according to these discussions. More critical incidents were added to fill in the gaps. At this stage, participants were asked to reassign all critical incidents to their dimensions and to rate them on a 9-point scale. Approximately 30% of the critical incidents were eliminated. The completed rating scales for the nine identified dimensions consisted of a definition and a 9-point continuum described with specific behavioral incidents. The second rating method was developed by breaking down the definitions generated in the workshops into their major elements. These statements were used as Likert-type items with a 4-point response format. Results showed high convergent validity ( $ICC = .481$ ,  $p < .001$ ) and low discriminant validity ( $ICC = .140$ ,  $p < .001$ ). There was also low method bias ( $ICC = .113$ ,  $p < .001$ ) and moderate error variance ( $VC = .29$ ). (7 references)

Dickinson, T. L., & Tice, T. E. (1973). A multitrait-multimethod analysis of scales developed by retranslation. Organizational Behavior and Human Performance, 9, 421-432.

One hundred forty-nine firefighters from four municipal fire departments in the midwest were rated by their immediate superiors and by a peer of their choice. All raters participated in the development of performance dimensions and behavioral statements in a series of workshops. The procedure resulted in 40 statements that illustrated three dimensions. The statements were used as a 40-item checklist. An analysis of the multitrait-multimethod correlation matrix indicated that there was low convergent validity ( $ICC = .179$ ,  $p < .001$ ) and low discriminant validity ( $ICC = .072$ ,  $p < .001$ ). The data also indicated that there was also a moderate degree of method bias ( $ICC = .273$ ,  $p < .001$ ) and high error variance ( $VC = .54$ ). (21 references)

Finley, D. M., Osburn, H. G., Dubin, J. A., & Jeanneret, P. R. (1977). Behaviorally based rating scales: Effects of specific anchors and disguised continua. Personnel Psychology, 30, 659-669.

Study 1. Sixty female managers of retail department stores located in small towns were evaluated on their job performance by eight district supervisors (first-line) and five regional supervisors (second-line). A behaviorally general, mixed standard scale was used. This scale consisted of three statements for each of 12 performance dimensions. The statements described highly effective performance, average performance, and highly ineffective performance. Raters were asked to indicate whether the ratee was better than, the same as, or worse than each statement. The statements were presented in a random order. Results indicated high convergent validity ( $ICC = .325$ ,  $p < .001$ ) and low discriminant validity ( $ICC = .088$ ,  $p < .001$ ). There was also moderate method bias ( $ICC = .202$ ,  $p < .001$ ) and high error variance ( $VC = .44$ ).

Study 2. On a second occasion, the same ratees described above were rated by the same raters described above using a behaviorally general, behaviorally anchored rating scale (BARS). This

scale consisted of three statements of highly effective, average, and highly ineffective performance for each of 12 performance dimensions. These statements were placed at points 6, 4, and 2 on a 7-point continuum. Results indicated high convergent validity ( $ICC = .438, p < .001$ ) and low discriminant validity ( $ICC = .105, p < .001$ ) and method bias ( $ICC = .124, p < .001$ ). There was also high error variance ( $VC = .35$ ).

Study 3. Fifty-three department store managers were rated by seven first-line supervisors and by five second-line supervisors. The ratees belonged to the same organizations as those in the previous two studies. The behaviorally general, BARS described in Study 2 was used to collect the ratings. The analyses revealed high convergent validity ( $ICC = .380, p < .001$ ), low discriminant validity ( $ICC = .087, p < .001$ ), and low method bias ( $ICC = .174, p < .001$ ). There was also high error variance ( $VC = .33$ ).

Study 4. On a second occasion, the same raters as in Study 3 rated the same ratees as in Study 3, using a behaviorally specific, BARS. This rating scale was developed following behavioral expectation scale procedures. Specific behavioral anchors were placed along a 7-point continuous scale for each of the 12 dimensions. Analysis of the multitrait-multimethod correlation matrix showed high convergent validity ( $ICC = .532, p < .001$ ), low discriminant validity ( $ICC = .037, p < .005$ ), and a low degree of method bias ( $ICC = .119, p < .001$ ). Error variance was high ( $VC = .34$ ).

Study 5. Sixty-four department store managers were evaluated by eight first-line supervisors and five second-line supervisors, using the behaviorally general, mixed standard scale described in Study 1. Results indicated moderate convergent validity ( $ICC = .283, p < .001$ ) and low discriminant validity ( $ICC = .062, p < .005$ ). There was also low method bias ( $ICC = .195, p < .001$ ) and high error variance ( $VC = .52$ ).

Study 6. The same raters and ratees as in Study 5 were used in this study. Ratings were made 9 weeks apart. The rating scale used in this study was the behaviorally specific scale described in

Study 4. The results showed high convergent validity ( $ICC = .487, p < .001$ ), low discriminant validity ( $ICC = .035, p < .005$ ), and low method bias ( $ICC = .177, p < .001$ ). Error variance was high ( $VC = .34$ ). (6 references)

Gunderson, E. K., & Ryman, D. H. (1971). Convergent and discriminant validities of performance evaluations in extremely isolated groups. Personnel Psychology, 24, 715-724.

One hundred five civilian scientists and Navy personnel participated in this study. All ratees had spent the winter at one of five Antarctic stations during a 4-year period. Two station leaders, the officer in charge, and the scientific leader rated each station member on three dimensions of performance. Station members were also asked to make three to five nominations (depending on the size of the station membership) for each item. Thus, there were supervisor ratings and peer nominations for the three dimensions. The results indicated a high degree of convergent validity ( $ICC = .449, p < .001$ ) and low discriminant validity ( $ICC = .168, p < .001$ ). There was also low method bias ( $ICC = .107, p < .001$ ) and high error variance ( $VC = .31$ ). (6 references)

Heneman, H. G., III. (1974). Comparisons of self- and superior ratings of managerial performance. Journal of Applied Psychology, 59, 638-642,

Ratings were received from 102 master of business administration graduates and their immediate supervisors. A questionnaire was mailed to the participants where they were currently employed. Instructions assured the participants that their responses would remain confidential and would be used for research purposes only. The rating form consisted of eight dimensions. A brief behavioral description of each dimension and a 7-point rating scale (from low performance to high performance) were placed on the performance evaluation form. Results indicated that there was little evidence of discriminant validity ( $ICC = .098, p < .002$ ) and only moderate support for convergent validity ( $ICC = .202, p < .001$ ). There was also little evidence of method bias ( $ICC = .190, p < .001$ ) with high error variance ( $VC = .58$ ). (15 references)

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. Journal of Applied Psychology, 63, 579-588.

Study 1. Ninety-seven management employees of a medium-sized manufacturing concern participated in this study. These participants included all levels of management. The rating instrument consisted of seven items covering seven dimensions of work performance (overall performance was eliminated from the analyses). All participants were assured that their responses would be used for research purposes only and would remain confidential. Each manager was asked to evaluate his own performance, the performance of a specified peer, and, where appropriate, the performance of direct managerial and professional subordinates. Results indicated moderate support for convergent validity ( $ICC = .249$ ,  $p < .001$ ). However, there was little evidence of discriminant validity ( $ICC = .068$ ,  $p < .001$ ) with a high degree of method bias ( $ICC = .395$ ,  $p < .001$ ). There was also high error variance ( $VC = .34$ ).

Study 2. As part of the study described above, 64 professional employees of the same organization were also evaluated. These employees included engineering, marketing, and other specialized personnel. The same rating instrument and rating procedures described above were employed for this sample of ratees. Analysis of the multitrait-multimethod correlation matrix for these employees also indicated a moderate level of convergent validity ( $ICC = .232$ ,  $p < .001$ ), high method bias ( $ICC = .393$ ,  $p < .001$ ), and little discriminant validity ( $ICC = .054$ ,  $p < .001$ ). There was also high error variance ( $VC = .38$ ). (19 references)

Ivancevich, J., M. (1977). A multitrait-multirater analysis of a behaviorally anchored rating scale for sales personnel. Applied Psychological Measurement, 1, 523-531.

Eight regional sales managers and 14 district sales managers evaluated 102 sales personnel of a large national organization. Behaviorally anchored rating scales were developed in five phases: (a) six randomly selected district sales managers and sales people identified independent job performance

dimensions and defined each with three general critical incident statements; (b) a second group of 12 reviewed the results of Phase I and wrote three specific critical incidents for those dimensions retained; (c) a new group discussed the results of Phases I and II and assigned the list of critical incidents to the dimensions, retaining those incidents for which there was 70% agreement; (d) another group rated the descriptors on a scale with intervals of .25 ranging from .00 to 2.00; and (e) the final raters were trained on the use and development of the rating form and common performance appraisal errors. Sales personnel were eventually rated on six performance dimensions. Results indicated high convergent validity ( $ICC = .325$ ,  $p < .001$ ), low discriminant validity ( $ICC = .171$ ,  $p < .001$ ), low method bias ( $ICC = .108$ ,  $p < .001$ ), and high error variance ( $VC = .44$ ). (17 references)

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analysis of ratings. Psychological Bulletin, 75, 34-49.

As part of a larger study, each of 183 department heads at the Owens Illinois Company was rated by his/her superior (plant manager), his/her most effective subordinate (foreman), and his/her least effective subordinate (foreman). Department heads were evaluated on 20 performance dimensions which were derived from existing scales. A detailed description of each dimension accompanied the rating scale. The data showed moderate convergent validity ( $ICC = .268$ ,  $p < .001$ ), low discriminant validity ( $ICC = .051$ ,  $p < .001$ ), and a high degree of method bias ( $ICC = .342$ ,  $p < .001$ ). There was also high error variance ( $VC = .39$ ). (37 references)

Lee, R., Malone, M., & Greco, S. (1981). Multitrait-multimethod-multirater analysis of performance ratings for law enforcement personnel. Journal of Applied Psychology, 66, 625-632.

In this study, 144 deputy sheriffs from three counties in Wisconsin were rated by the two supervisors most familiar with their job performance. Raters rated each ratee with two appraisal forms: (a) a summated rating scale, and (b) a graphic rating scale. All raters received

intensive training in a small group setting. The rating forms and instructions were presented and discussed. Common rating errors were also discussed. All raters were assured that their ratings would be used for test validation purposes only. The results indicated that the rating scales possessed high convergent validity ( $ICC = .563$ ,  $p < .001$ ), moderate discriminant validity ( $ICC = .262$ ,  $p < .001$ ), and no method bias ( $ICC = .000$ ,  $p > .05$ ). There was low error variance ( $VC = .18$ ). (11 references)

Marsh, H. (1982). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. Journal of Educational Psychology, 74, 264-279.

Students in 329 courses in the Division of Social Sciences at the University of Southern California evaluated their instructors. The students were told that their evaluations would be used to provide feedback to instructors and would be considered as part of personnel decisions. The evaluation instrument consisted of 35 items covering nine dimensions. The instructors of the 329 courses rated themselves. The participation of the instructors was voluntary and all were guaranteed confidentiality. The results showed low convergent validity ( $ICC = .129$ ,  $p < .001$ ) and high discriminant validity ( $ICC = .301$ ,  $p < .001$ ). There was low method bias ( $ICC = .151$ ,  $p < .001$ ), as well as high error variance ( $VC = .47$ ). (28 references)

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. Journal of Educational Psychology, 71, 149-160.

Student evaluations were collected for 83 undergraduate courses taught by the faculty in the Division of Social Sciences at the University of Southern California. The instructors of these courses also evaluated themselves, using the same rating form as the students except that the items were worded in the first person. Thirty-two of the instructors evaluated two courses they deemed to be "most effective" and "least effective." The remaining 19 instructors evaluated only one course.

All faculty members were assured that their responses would remain confidential. The evaluation form consisted of 24 items covering six dimensions or traits. There was little support for convergent validity ( $ICC = .179$ ,  $p < .001$ ), with moderate discriminant validity ( $ICC = .294$ ,  $p < .001$ ). There was also evidence that the ratings contained little method bias ( $ICC = .167$ ,  $p < .001$ ) and high error variance ( $VC = .42$ ). (25 references)

Nealey, S. M., & Owen, T. W. (1970). A multitrait-multimethod analysis of predictors and criteria of nursing performance. Organizational Behavior and Human Performance, 5, 348-365.

Five unit supervising nurses (first-level supervisors) and the head of the nursing service and his assistant (second-level supervisors) rated 25 head nurses in a Veterans Administration Hospital. The first-level supervisors rated only those nurses who were their immediate subordinates whereas the second-level supervisors rated all head nurses. There were three dimensions, and each dimension consisted of one 5-point scale which ranged from "much above average for the unit" to "a little below average for the unit." The results indicated a high level of convergent validity ( $ICC = .427$ ,  $p < .001$ ) and no discriminant validity ( $ICC = .000$ ,  $p < .819$ ). There was also moderate evidence of method bias ( $ICC = .265$ ,  $p < .001$ ), as well as high error variance ( $VC = .41$ ). (10 references)

Orpen, C. (1973). An empirical assessment of the job performance of high-level executives by means of a multitrait-multimethod matrix. Psychologia Africana, 15, 7-14.

Sixty-three South African business executives from various economic sectors were each rated by a superior, peer, and subordinate who were familiar with their work. Every ratee was rated on five dimensions, using a 5-point scale ranging from "always" to "never." Each dimension was accompanied by a carefully worked out definition. These definitions were based on capsule descriptions written by the raters in a pilot study. An analysis of the multitrait-multimethod correlation matrix indicated that there was high convergent validity ( $ICC = .322$ ,  $p < .001$ ), low discriminant validity

(ICC = .121,  $p < .001$ ), and low method bias (ICC = .044,  $p < .005$ ). There was also considerable error variance (VC = .54). (23 references)

Tucker, M. F., Cline, V. B., & Schmitt, J. R. (1967). Prediction of creativity and other performance measures from biographical information among pharmaceutical scientists. Journal of Applied Psychology, 51, 131-138.

Study 1. As part of a larger study, supervisory and peer ratings were obtained for 79 scientists employed in the scientific division of the Vicks Chemical Company, the Merrell Drug Company and the National Drug Company. All ratees were male and had obtained at least a BA or BS degree. All ratees were evaluated on three dimensions of on-the-job performance. Analysis of the multitrait-multimethod correlation matrix indicated high convergent validity (ICC = .355,  $p < .001$ ) and low discriminant validity (ICC = .049,  $p < .002$ ). There was also high method bias (ICC = .431,  $p < .001$ ), with moderate error variance (VC=.22).

Study 2. Ratings were obtained for a second set of 78 scientists. These scientists belonged to the organizations mentioned above and had the same characteristics as the ratees described in Study 1. Scientists were rated by their supervisors and peers on three dimensions of work performance. Results showed high convergent validity (ICC = .315,  $p < .001$ ) and low discriminant validity (ICC = .107,  $p < .001$ ). Again, however, there was high method bias (ICC = .448,  $p < .001$ ), along with low error variance (VC = .18). (10 references)

Zedeck, S., & Baker, H. T (1972). Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. Organizational Behavior and Human Performance, 7, 457-466.

Nine head nurses and five supervisors rated 71 registered nurses in a public, non-profit hospital in Northern California. The head nurse and supervisor pair did not necessarily have a common set of registered nurses to evaluate. Behavioral expectation scales from a previous study were used to obtain the ratings. The raters received a brief training session on the use of the scales. Each

rater was asked to record one to five incidents of past ratee performance for each of five rating dimensions. The results indicated a high degree of convergent validity ( $ICC = .396, p < .001$ ) and low discriminant validity ( $ICC = .075, p < .001$ ). There was also moderate evidence of method bias ( $ICC = .247, p < .001$ ), as well as high error variance ( $VC = .33$ ).  
(9 references)

APPENDIX B: CODE SHEET WITH STUDY FREQUENCIES FOR  
RESPONSE CATEGORIES AND MEANS FOR CONTINUOUS ITEMS

An asterisk (\*) on an item indicates that a study may be coded in more than one category on that item.

IDENTIFICATION

1. Study ID #: N/A
2. Coder ID #: N/A
3. Source of study:  
N/A Book N/A Dissertation  
N/A Journal N/A Paper Presentation  
N/A Technical Paper N/A Unpublished Manuscript  
N/A Other \_\_\_\_\_  
(Describe)
4. Year study was published or written: N/A

TRAITS

- \*5. Which of the following procedures or techniques were used to collect the data for the development of the performance dimensions?
- 1 Job descriptions
- 1 Surveys (e.g., critical incident questionnaires)
- 14 Discussions with subject-matter experts (e.g., conferences, workshops)
- 11 The scale was derived from an existing scale
- 0 Other. \_\_\_\_\_  
(Describe)
- 7 Not stated

\*6. What procedures were used to derive the performance dimensions on which ratings are obtained?

7 Factor analysis

12 Retranslation: Smith and Kendall (1963)

5 Expert prescriptions

0 Other \_\_\_\_\_  
(Describe)

12 Not stated

\*7. Which best describes the content of the performance dimensions?

26 Behavioral; job tasks or activities such as communicating or planning

11 Trait; human attributes such as drive, effort, or initiative

6 Other \_\_\_\_\_  
(Describe)

0 Not stated

8. Which best describes the specificity of the content of the performance dimensions?

21 Specific; detailed definitions of the dimension; examples may be provided to define levels of the dimension

8 General; only the titles or brief definitions of dimensions; brief adjectives may be provided to define levels of the dimension

2 Not stated

9. How many distinct dimensions were defined for the multitrait-multimethod matrix? 7.16 (Number)

10. How many separate ratings determined a ratee's dimension score? 2.06 (Average) Information available from 21 studies.

11. What was the mean test-retest reliability reported for the performance dimensions? .67 (Average) Information available from 3 studies.
- \*12. What were the raters told was the purpose for obtaining ratings?
- 1 Criterion-related validation
- 4 Basic research
- 2 Employee growth and development
- 1 Administrative decisions
- 0 Other \_\_\_\_\_ (Describe)
- 24 Not stated
13. Was the rating format the "method" in the MTMM design?
- 3 Yes 28 No
- 13a. If "Yes": To what degree was rater and ratee variance confounded in the individual ratings of ratee performance? Check the statement which best describes the study's design.
- Format A \_\_\_\_\_ (Identify)
- 0 Crossed procedure: All raters rated all ratees.
- 0 Nested procedure: A different set of raters rated each ratee.
- 3 Mixed procedure: Some raters rated several but not all ratees.
- Format B \_\_\_\_\_ (Identify)
- 0 Crossed procedure: All raters rated all ratees.

0 Nested procedure: A different set of  
raters rated each ratee.

3 Mixed procedure: Some raters rated  
several but not all ratees.

Format C \_\_\_\_\_  
(Identify)

0 Crossed procedure: All raters rated all  
ratees.

0 Nested procedure: A different set of  
raters rated each ratee.

0 Mixed procedure: Some raters rated  
several but not all ratees.

14. Was rating source the "method" in the MTMM design?

28 Yes 3 No

14a. If "Yes": To what degree was rater and ratee  
variance confounded in the individual ratings of  
ratee performance? Check the statement which  
best describes the study's design.

Source A \_\_\_\_\_  
(Identify)

0 Crossed procedure: All raters rated all  
ratees.

8 Nested procedure: A different set of  
raters rated each ratee.

20 Mixed procedure: Some raters rated  
several but not all ratees.

Source B \_\_\_\_\_  
(Identify)

0 Crossed procedure: All raters rated all  
ratees.

7 Nested procedure: A different set of  
raters rated each ratee.

20 Mixed procedure: Some raters rated several but not all ratees.

Source C \_\_\_\_\_  
(Identify)

0 Crossed procedure: All raters rated all ratees.

6 Nested procedure: A different set of raters rated each ratee.

2 Mixed procedure: Some raters rated several but not all ratees.

15. Was rater training provided?

3 Yes 28 No

\*15a. If "Yes": What was the nature of the rater training?

1 Group discussion

0 Videotape

1 Lecture

0 Written instructions on administration/use of the rating form

0 Other \_\_\_\_\_  
(Describe)

2 Not stated

\*15b. What was the stated focus of the rater training program?

2 Reduce psychometric errors

0 Improve accuracy of ratings

3 Impart knowledge of rating procedures

0 Other \_\_\_\_\_  
(Describe)

0 Not stated

16. In what setting was the study conducted?

- 0 Laboratory; observing actual videotapes of  
ratee performance
- 0 Laboratory; observing actual performance  
by ratees
- 27 Organizational/field
- 4 Classroom
- 0 Other \_\_\_\_\_  
(Describe)
- 0 Not stated

#### METHODS

\*17. What were the rating formats?

- 5 BARS: No expectation terminology
- 3 BES: Expectation terminology
- 0 BOS: Summated scaling with agree/disagree  
anchors
- 2 Mixed Standard: +, 0, - responses
- 15 Graphic: Numerical and/or adjectival  
anchors
- 6 Other \_\_\_\_\_  
(Describe)
- 4 Not stated

\*18. Who participated in the development of the rating  
scales?

- 6 Raters
- 4 Ratees
- 6 Experts (e.g., psychologists, consultants,  
or subject-matter experts)
- 4 The scale was an existing scale  
(i.e., taken off the shelf and used unchanged)

7 The scale was a modified version of an existing scale

1 Other \_\_\_\_\_  
(Describe)

14 Not stated

#### RATER CHARACTERISTICS

\*19. Who were the sources for ratings?

25 1st level supervisors

12 2nd level supervisors

10 Peers

10 Self

4 Subordinates

4 Students

0 Other \_\_\_\_\_  
(Describe)

0 Not stated

\*20. What was the raters' sex?

12 Male

3 Female

19 Not stated

#### RATEE CHARACTERISTICS

\*21. What was the sex of the ratees?

4 Male

7 Female

21 Not stated

22. What was the number of ratees employed in this study (i.e., the number of instructors, workers, videotapes, or paper people that were rated)?  
104.52 (Number)

23. What was the number of ratees per rater?  
6.87 (Average) Information available from  
14 studies.

24. Who were the ratees?

0 Videotapes of people performing work

31 People in a work setting

0 Other \_\_\_\_\_  
(Describe)

\*25. What was the organizational affiliation of the ratees?

17 Private industry

2 Military

6 Academia

6 Public sector organization (nonprofit,  
government related)

0 Other \_\_\_\_\_  
(Describe)

1 Not stated

26. What was the nature of the work performed by the ratees?

N/A  
(Title and nature of the duties performed)

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

0 Not stated

## APPENDIX C: CODE BOOK

An asterisk (\*) on an item indicates that a study may be coded in more than one category on that item.

### IDENTIFICATION

#### 1. Study ID #.

This will appear in the upper right-hand corner on the first page of each study.

NOTE: Some studies will be coded more than once because more than one multitrait-multimethod matrix will be reported. In these cases the Study ID # will have letter subscripts and, in parentheses, the study table number of the multitrait-multimethod matrix will be identified. Both the Study ID # and the letter subscript should be recorded. All studies with letter subscripts should be coded separately for each subscript.

#### 2. Coder ID #.

Terry Dickinson----	1
Catherine Hassett--	2
Scott Tannenbaum---	3

#### 3. Source of study.

Check the source from which the study was obtained.

#### 4. Year study was published or written.

This should be found on the first page of the study. Record only the last two digits.

### TRAITS

#### \*5. Which of the following procedures or techniques were used to define work content?

Do not confuse these procedures with those used to cluster work content to derive performance dimensions.

NOTE 1: More than one category may be appropriate when coding this item. For example, both job descriptions and surveys could have been used to define work content.

NOTE 2: Use of the Smith and Kendall (1963) procedure entails discussions with subject-matter experts.

- \*6. What procedures were used to derive the performance dimensions on which ratings are obtained?

This refers to the clustering of work content in order to derive job performance dimensions.

NOTE: Although more than one check may be appropriate, this will be the exception rather than the rule.

- \*7. Which best describes the content of the performance dimensions?

Content is determined by the definition of a dimension and its anchors or examples. Refer to a dimension's title only when other information is unavailable.

Behavior-oriented indicates job tasks or activities necessary to perform the work. Tasks or activities would include planning, communicating, typing, relating to customers, etc.

Trait-oriented indicates attributes of the person who is performing the job. These traits would include drive, knowledge, motivation, effort, initiative, etc.

NOTE: Both categories would be checked only if a set of dimensions is clearly trait-oriented and another set is behavior-oriented.

8. Which best describes the specificity of the content of the performance dimensions?

Specific content indicates that detailed definitions of the dimensions are given. Examples may also be provided to define low, medium, and high levels of the dimension scales.

General content indicates that only the titles or brief definitions of the dimensions are given. Brief adjectives may be provided to define low, medium, and high levels of the dimension scales.

9. How many distinct dimensions were defined for the multitrait-multimethod matrix?

This refers to the number of dimensions and not the number of items on a checklist or questionnaire that was used to measure the dimensions. Remember that each dimension is operationally defined by two or more formats or sources. Count each dimension only once.

NOTE: Do not include a global performance dimension as part of the matrix when specific dimensions have been defined. Include a global performance dimension only when other global dimensions such as work effort and job commitment are defined for the matrix.

10. How many separate ratings determined a ratee's dimension score?

This refers to the number of ratings made by a single rater to determine a ratee's dimension score. For a behaviorally anchored rating scale, this number would equal 1. For a behavioral observation scale, the number would equal the number of items on the dimension's scale. For a mixed standard scale, the number would equal 3. If the number of ratings varies across dimensions, provide the average (calculate all averages to two significant decimal places). If no information is available, record a 0 in the space provided.

11. What was the mean test-retest reliability reported for the performance dimensions?

If test-retest reliability is not reported, record a 0 in the space provided. This will serve as a missing value.

- \*12. What were raters told was the purpose for obtaining ratings?

Criterion-related validation and basic research should be distinguished in the study. However, if

the study states only that the ratings were obtained for research purposes, check the purpose as basic research.

Employee growth and development as a purpose includes feedback to the ratees.

Administrative decisions as a purpose include hiring, firing, promotion, transfer, and pay increases.

NOTE: More than one check may be appropriate where different raters were rating for different purposes.

13. Was rating format the "method" in the MTMM design?

Check "Yes" if different rating formats were used as methods in the MTMM matrix. If "Yes," answer 13a. If "No," go to item 14.

- 13a. For each format, identify the procedure that was used to collect ratings. Example procedures are displayed below.

Crossed procedure: All raters rated by all ratees.

Format A e.g., BARS  
(Identify)

Ratee 1 by Rater 1  
Rater 2  
Rater 3

Ratee 2 by Rater 1  
Rater 2  
Rater 3

Nested procedure: A different set of raters rated each ratee.

Format B e.g., BOS  
(Identify)

Ratee 1 by Rater 1  
Ratee 2 by Rater 2  
Ratee 3 by Rater 3

**OR**

Format B e.g., MSS  
(Identify)

**Rater 1 by Rater 1                  to calculate  $\bar{X}$  as the**

**Rater 2                  rating**

Ratee 2 by Rater 3  
Rater 4 to calculate  $\bar{X}$  as the  
Rater 5 rating

**Mixed procedure:** Some raters rated several but not all ratees.

Format C e.g., 0  
(Identify)

Ratee 1	by	Rater 1
Ratee 2	by	Rater 1
Ratee 3	by	Rater 2
Ratee 4	by	Rater 3
Ratee 5	by	Rater 4
Ratee 6	by	Rater 4

**NOTE: If less than three formats were used, record 0 as the identification for "Format C".**

14. Was rating source the "method" in the MTMM design?

Check "Yes" if the raters were used as methods in the MTMM matrix. If "Yes," answer 14a. If "No," go to item 15.

- 14a. For each source, identify the procedure that was used to collect ratings. Example procedures are displayed below.

**Crossed procedure:** A rating entity rated all ratees.

Source A e.g., The supervisor  
(Identify)

Ratee 1  
Ratee 2  
Ratee 3  
.  
.  
.  
Ratee n

OR

Source A e.g., (n-1) Peers  
(Identify)

Ratee 1  
Ratee 2  
Ratee 3  
.  
.  
.  
Ratee n

Nested procedure: Different source entities rated  
each ratee.

Source B e.g., The Supervisor  
(Identify)

Ratee 1 by Supervisor 1  
Ratee 2 by Supervisor 2  
Ratee 3 by Supervisor 3

OR

Source B e.g., Peers  
(Identify)

Ratee 1 by Peer 1  
Peer 2 to calculate  $\bar{X}$  as the rating

Ratee 2 by Peer 3  
Peer 4 to calculate  $\bar{X}$  as the rating

Ratee 3 by Peer 5  
Peer 6 to calculate  $\bar{X}$  as the rating  
Peer 7

Mixed procedure: Some source entities rated  
several but not all raters.

Source C e.g., Supervisors  
(Identify)

Ratee 1 by Rater 1  
Ratee 2 by Rater 1  
Ratee 3 by Rater 2  
Ratee 4 by Rater 3  
Ratee 5 by Rater 4  
Ratee 6 by Rater 5

OR

Source C e.g., Peers  
(Identify)

Ratee 1 by Peer 1  
Peer 2 to calculate  $\bar{X}$  as the rating

Ratee 2 by Peer 2

Ratee 3 by Peer 2  
Peer 3 to calculate  $\bar{X}$  as the rating

NOTE: If less than three sources were used,  
record 0 as the identification for "Source C".

15. Was rater training provided?

If rater training was provided, it should be  
stated explicitly in the article. If not stated,  
check "No."

15a. If "Yes": What was the nature of the rater  
training?

Group discussion indicates that the raters  
discussed rating strategies, rating errors,  
dimension or item meanings, or rating methods.

Videotape indicates that the raters watched a  
videotape explaining the above.

Lecture indicates that the raters listened to an  
expert explain the above in-person.

Written instructions indicates that the raters  
received written instructions on how to use and  
administer the rating form.

AD-A174 759

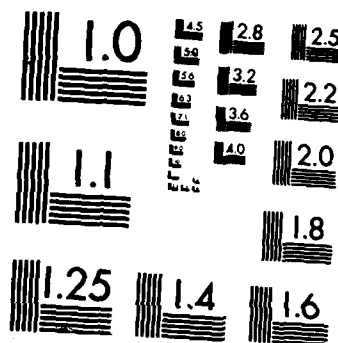
WORK PERFORMANCE RATINGS: A META-ANALYSIS OF  
MULTITRAIT-MULTIMETHOD STUDIES(U) TEXAS MAXIMA CORP SAN  
ANTONIO T L DICKINSON ET AL DEC 86 AFHRL-TP-86-32  
F33615-83-C-0030 F/G 5/9

2/2

UNCLASSIFIED

NL





MICROCOPY RESOLUTION TEST CHART  
NATIONAL BUREAU OF STANDARDS-1963-A

NOTE: More than one check may be appropriate for this item.

\*15b. What was the focus of the rater training program?

Reduce psychometric errors. This indicates that the training program was aimed at reducing psychometric errors such as leniency/severity or halo.

Improve accuracy. This indicates that the goal of training was to improve the raters' accuracy using such measures as differential accuracy, elevation, stereotype accuracy, or differential elevation.

Impart knowledge. This indicates that the focus of rater training was simply to teach raters the correct use of the rating form.

NOTE: More than one check may be appropriate for this item.

16. In what setting was the study conducted?

This should be specifically stated in the article. If it is not, check "Not stated."

METHODS

\*17. What were the rating formats?

This refers to the manner in which the scales are displayed.

BARS. This refers to behaviorally anchored rating scales with no expectation terminology.

BES. This refers to behaviorally anchored rating scales using expectation terminology.

BOS. This refers to scales that were developed with summated scaling and which have almost always/almost never anchors. BOS have several items per dimension, and each of those items is rated on a Likert-type format.

Mixed Standard. This refers to a scale which indicates that three statements were given for each dimension and that raters responded to whether the ratee performed at that level

(0), above the level (+), or below the level (-).

Graphic. This refers to scales that have numerical anchors and adjectival anchors such as high, average, or low.

NOTE: - More than one check may be appropriate for this item.

- \*18. Who participated in the development of the rating scales?

It should state in the article how the scales were developed and who aided in this development.

"Raters" and "Ratees" should be checked only if those raters or ratees using the scale in the study were involved in the development of the scale. In cases where the scale is simply described, with no indication as to how it was developed, check "Not stated."

NOTE: More than one response may be appropriate for this item (e.g., the scale may have been a modified version of an existing scale and raters may have participated in the modification). In this case, check both "The scale was a modified version of an existing scale" and "Raters." An existing scale which remains virtually unchanged after discussion with raters/ratees should not be considered modified and only "existing scale" should be checked.

#### RATER CHARACTERISTICS

- \*19. Who were the sources for ratings?

The article should state who made the evaluations. If the study states that a ratee's boss or supervisor provided the ratings, without giving the specific level of either, check "1st level supervisor."

NOTE: More than one check may be appropriate for this item.

- \*20. What was the raters' sex?

NOTE: More than one check may be appropriate for this item.

## RATEE CHARACTERISTICS

- \*21. What was the sex of the ratees?

NOTE: More than one check may be appropriate for this item.

22. What was the number of ratees employed in this study?

This refers to the number of instructors, workers, videotapes, or paper people that were rated.

NOTE: Paper people are ratees whose performance is described on paper.

23. What was the number of ratees per rater?

How many ratees did each rater evaluate? This value will be 1 or greater. Sometimes not all the raters in a study will evaluate the same number of ratees. In these cases, report the mean number of ratees per rater by dividing the number of ratees by the number of raters. If more than one source provided ratings, compute a separate average for each source and then compute an average of the averages. If the study did not provide sufficient information to compute the average, record a 0 in this space.

24. Who were the ratees?

In some studies the ratees may be people in a work setting or videotapes of people performing work. In other studies, the ratees may be of some other form. For example, "Other" would include paper people.

- \*25. What was the organizational affiliation of the ratees?

The article should state the type of organization to which the ratees belong.

NOTE: Public sector organizations would include nonprofit organizations such as hospitals (nonprofit), United Way, drug referral hot lines, and local, state, or Federal government bodies such as police departments, fire departments, etc.

26. What was the nature of the work performed by the ratees?

If the job title of the ratee is given, write this in the space provided. Also write the nature of duties performed. Please be clear and succinct.

END

1-87

DT/C